Multilingual substitutes in WSI

Князьковой Виктории Игоревны

https://github.com/vknyazkova/Multilinguial_Substitutes_WSI

Table of contents

Word Sense Induction and Lexical Substitution approach

Word occurrences: → **Clustering:**

cell.n.74: *It's in **cell** phones and laptop computers.*
cell.n.76: *Whether it is **cell** phones, digital cameras or flat TV screens, Japan continues to dominate many sectors of technology…*
cell.n.75: *From the moment you pick up a lab report stating that the doctor has diagnosed cancerous **cells**, life starts to get tough*
cell.n.78: *The free radicals are released when oxygen carrying red blood **cells** called haemoglobin die*

cell.n.74
cell.n.76

cell.n.75
cell.n.78

Word Sense Induction and Lexical Substitution approach

Word substitutes:

Clustering:

cell.n.74: *It's in **[phone cellphone phone-like]** phones and laptop computers.*
cell.n.76: *Whether it is **[cellphone phone phone-like]** phones, digital cameras or flat TV screens, Japan continues to dominate many sectors of technology...*
cell.n.75: *From the moment you pick up a lab report stating that the doctor has diagnosed cancerous **[tumor carcinoma cancer-derived]**, life starts to get tough*
cell.n.78: *The free radicals are released when oxygen carrying red blood **[blood-cell red-blood lymphocyte]** called haemoglobin die*

cell.n.74
cell.n.76

cell.n.75
cell.n.78

Experimental setup: Dataset

<cell.n.16>The team used a battery of the newly developed `` gene probes , '' snippets of genetic material that can track a gene 's presence in a cell . <TargetSentence>By analyzing cells extracted from eye tumors , they found defects in the second copy of chromosome 13 in the exact area as in the first copy of the chromosome . </TargetSentence>The finding riveted medicine . </cell.n.16>

Semeval 2010 Task 14

- 100 words
- 50 nouns
- 50 verbs
- Up to 3 context sentences

Experimental setup: Language models + Alignment

Fasttext language models[1]

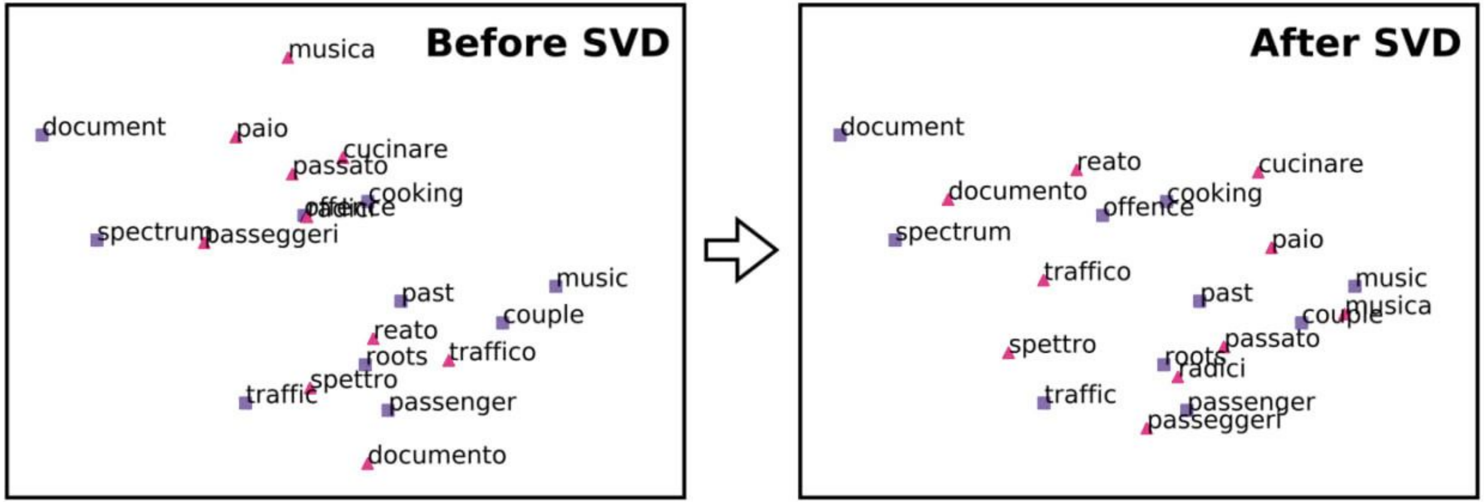
en, ru, fr, es, de

Alignment matrices[2]

Mapps each vector into shared vector space

[1] https://fasttext.cc/docs/en/crawl-vectors.html
[2] https://github.com/babylonhealth/fastText_multilingual

Alignment



[1] https://fasttext.cc/docs/en/crawl-vectors.html
[2] https://github.com/babylonhealth/fastText_multilingual

Substitutes Generation

Target

access.n

Multilingual targets

access, доступ, acceso

k-nearest neighbours

acess access acces Access access.We access. accessing

Cleaned neighbours

accessing

Context extraction

news media had to confidential information related

Context embeddings

Sorted by closeness to context substitutes

information provide resources просочиться проникнуть демо-доступ

Context extraction

dummy

n words from both the left and right sides of the target word, excluding the word itself

pos excluding

Same, but exclude [’PUNCT’, ’DET’, ’PART’, ’X’, ’AUX’] pos-tags

Substitutes cleaning

1. Remove attached punctuation marks

access.We access.

access access

2. Remove typos using Levenshtein distance

- pairwise comparison
- if Levenshtein distance > threshold remove from pair the less frequent substitute

acces
Access → *Access*

Clustering

01 Vectorization

Vectorize substitutes for every instance of target

02 Clustering algorithms

Either Agglomerative clustering or K-means

03 Number of clusters

Different strategies for selecting number of clusters

Clustering: vectorization

tf-idf

Не учитывает ранг подстановки

access.n.1: ‘provide allow able’
access.n.2: ‘able get allow’

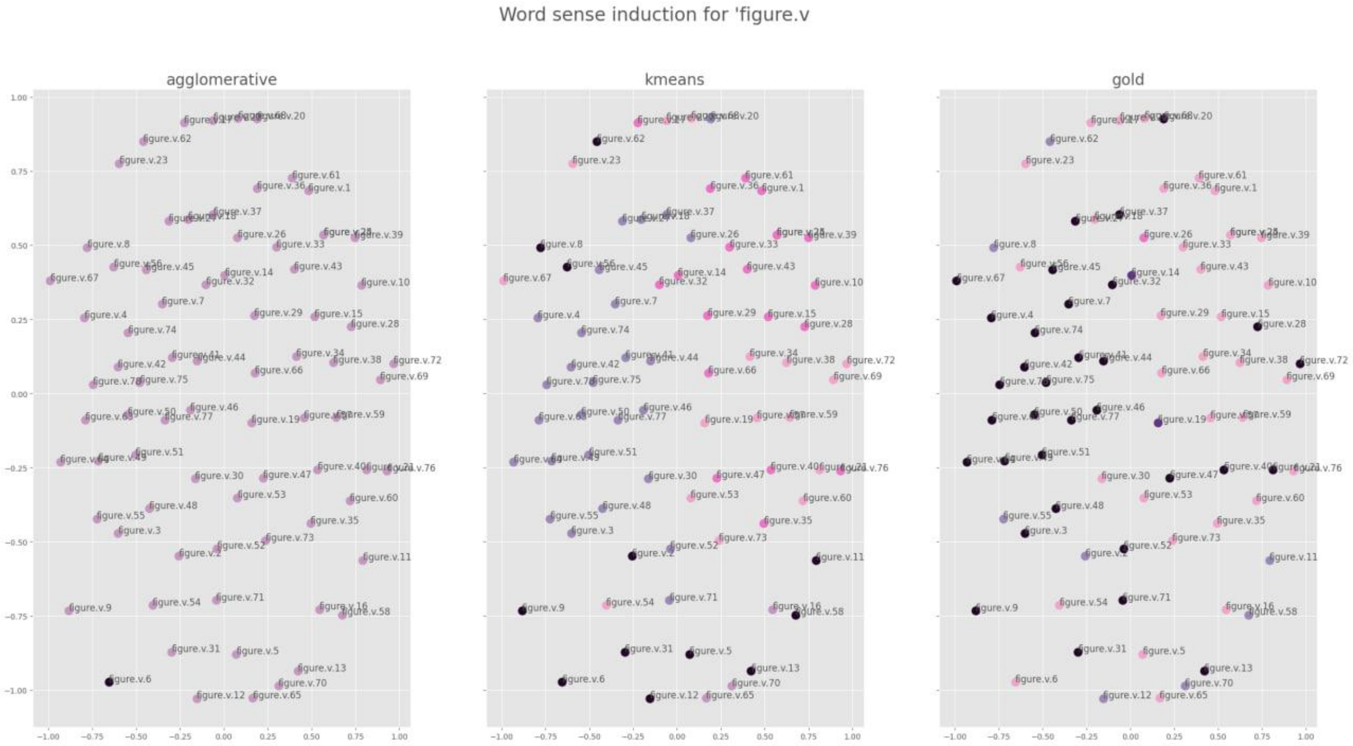
tf-idf-weighted

Вес подстановок зависит от ее ранга в списке замен

access.n.1: ‘provide provide provide allow allow able’
access.n.2: ‘able able able get get allow’

\+ *both methods **without** idf*

Clustering: algorithms



Agglomerative

faster
better overall results

but: *unbalanced clusters*

K-means

much slower
worse mean metrics

but: *more plausible clusters*

Clustering: number of clusters

fix

Predetermined number of clusters for all target words

Maximizing silhouette score

Selecting from some range number of clusters, that maximizes silhouette score

Predetermined range

Default: range(2, 10)

From 2 to number of contexts

For every target its own range

Experiments: english only

Best mean metric:** agglomerative with **tf-idf-weighted

lang	n subst	vecotrizer	clusterizer	context exctraction	fscore	precision	recall	vmeasure	homogenity	completeness	(fs * vm) ** 0.5
en	5	tf	agglomerative	dummy	51.816	71.888	47.930	12.892	21.405	14.484	20.719
en	5	tf	kmeans	dummy	38.580	37.312	50.360	16.968	15.525	22.761	22.248
en	5	tf-idf	agglomerative	dummy	40.870	47.004	49.574	18.343	20.059	25.958	23.180
en	5	tf-idf	kmeans	dummy	24.63	17.189	51.606	20.541	15.911	32.248	21.186
en	5	tf-weighted	agglomerative	dummy	48.490	64.675	49.121	14.836	22.006	18.861	22.241
en	5	tf-weighted	kmeans	dummy	38.775	36.476	52.373	17.936	16.179	23.866	23.702
en	5	tf-idf-weighted	agglomerative	dummy	43.593	51.530	49.506	18.537	21.765	24.116	24.388
en	5	tf-idf-weighted	kmeans	dummy	27.24	20.043	53.145	21.535	17.135	32.605	23.002
en	5	tf	agglomerative	pos_excl	47.959	63.83	48.076	13.822	19.432	17.353	20.904
en	5	tf	kmeans	pos_excl	36.510	34.894	50.528	15.862	14.412	21.785	20.977
en	5	tf-idf	agglomerative	pos_excl	38.068	41.439	50.491	19.311	19.069	27.720	23.724
en	5	tf-idf	kmeans	pos_excl	24.472	18.122	51.935	19.732	15.627	31.186	20.520
en	5	tf-weighted	agglomerative	pos_excl	44.489	54.053	48.656	15.967	19.083	19.936	22.719
en	5	tf-weighted	kmeans	pos_excl	33.428	29.026	50.563	17.671	14.952	25.067	21.548
en	5	tf-idf-weighted	agglomerative	pos_excl	38.187	39.925	50.081	19.634	18.676	28.423	23.857
en	5	tf-idf-weighted	kmeans	pos_excl	25.462	18.445	52.477	20.717	16.184	32.411	21.617

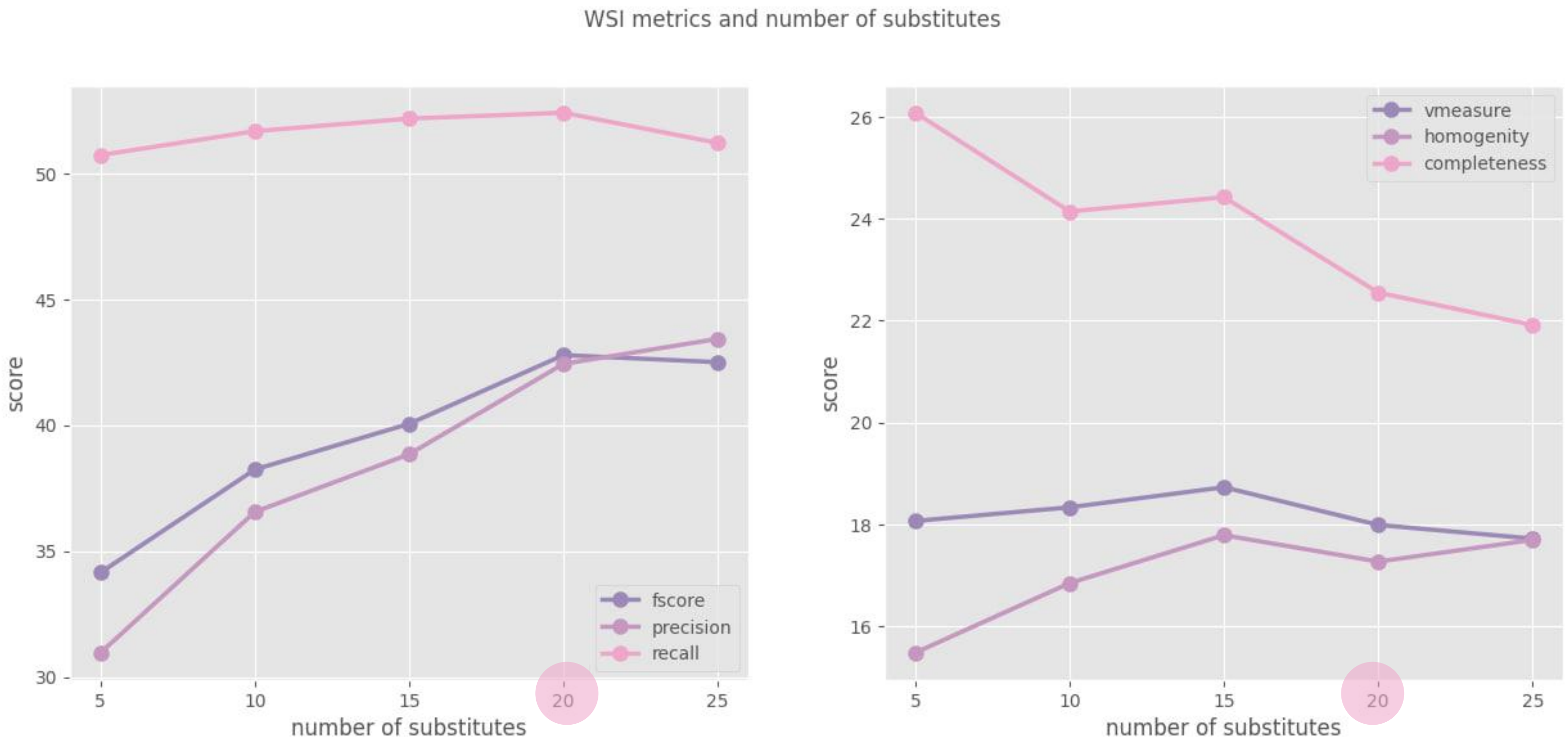
Experiments: english only

***Selected params:** kmeans with tf-idf-weighted*

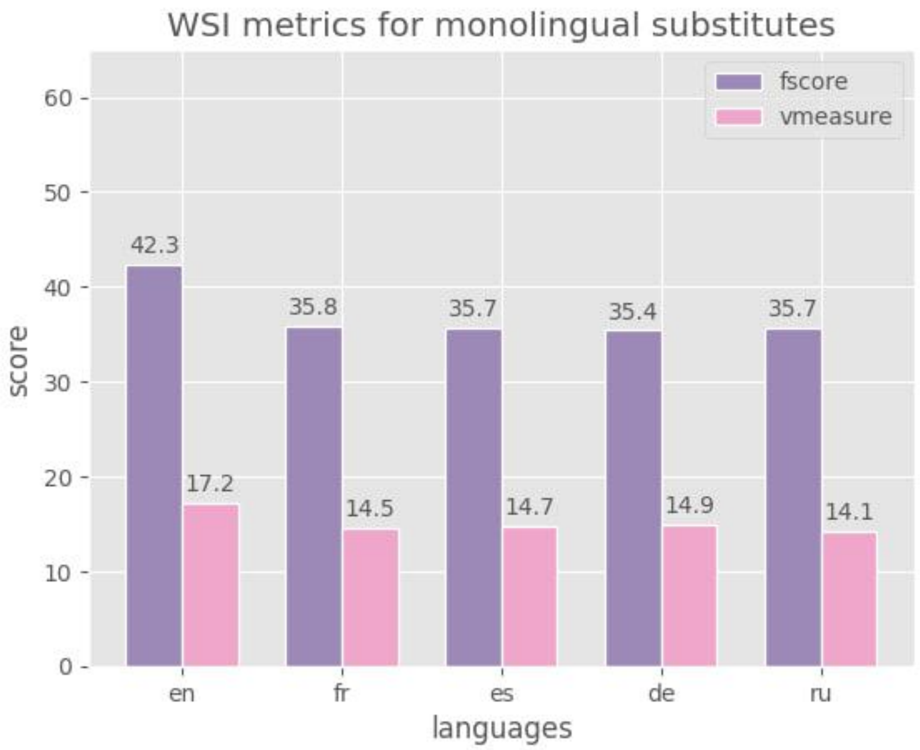
lang	n subst	vecotrizer	clusterizer	context exctraction	fscore	precision	recall	vmeasure	homogenity	completeness	(fs * vm) ** 0.5
en	5	tf	agglomerative	dummy	51.816	71.888	47.930	12.892	21.405	14.484	20.719
en	5	tf	kmeans	dummy	38.580	37.312	50.360	16.968	15.525	22.761	22.248
en	5	tf-idf	agglomerative	dummy	40.870	47.004	49.574	18.343	20.059	25.958	23.180
en	5	tf-idf	kmeans	dummy	24.63	17.189	51.606	20.541	15.911	32.248	21.186
en	5	tf-weighted	agglomerative	dummy	48.490	64.675	49.121	14.836	22.006	18.861	22.241
en	5	tf-weighted	kmeans	dummy	38.775	36.476	52.373	17.936	16.179	23.866	23.702
en	5	tf-idf-weighted	agglomerative	dummy	43.593	51.530	49.506	18.537	21.765	24.116	24.388
en	5	tf-idf-weighted	kmeans	dummy	27.24	20.043	53.145	21.535	17.135	32.605	23.002
en	5	tf	agglomerative	pos_excl	47.959	63.83	48.076	13.822	19.432	17.353	20.904
en	5	tf	kmeans	pos_excl	36.510	34.894	50.528	15.862	14.412	21.785	20.977
en	5	tf-idf	agglomerative	pos_excl	38.068	41.439	50.491	19.311	19.069	27.720	23.724
en	5	tf-idf	kmeans	pos_excl	24.472	18.122	51.935	19.732	15.627	31.186	20.520
en	5	tf-weighted	agglomerative	pos_excl	44.489	54.053	48.656	15.967	19.083	19.936	22.719
en	5	tf-weighted	kmeans	pos_excl	33.428	29.026	50.563	17.671	14.952	25.067	21.548
en	5	tf-idf-weighted	agglomerative	pos_excl	38.187	39.925	50.081	19.634	18.676	28.423	23.857
en	5	tf-idf-weighted	kmeans	pos_excl	25.462	18.445	52.477	20.717	16.184	32.411	21.617

Experiments: number of substitutes

***overall best:** 20 substitutes*

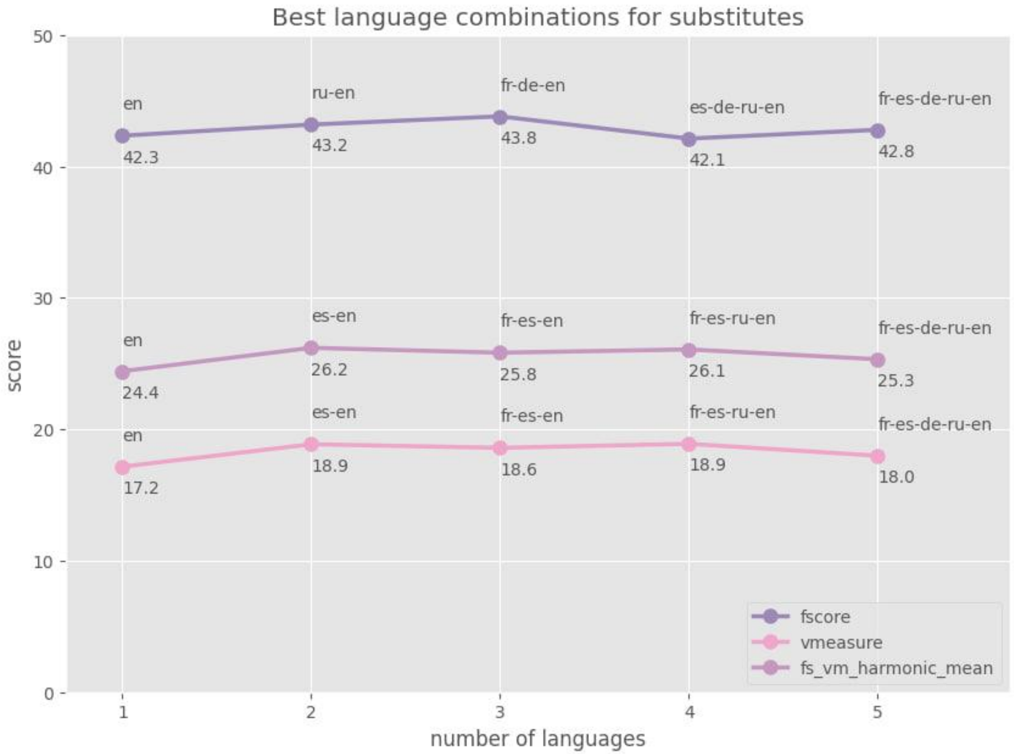


Experiments: Different languages



*English **outperforms** all other languages, when we use only one language for substitutes*

Results: Different combinations



*Incorporation of <u>additional languages</u> **can improve** english-only result (a little)*

*Some **language combinations** give better results than others*

-> for further studying

Conclusions

Multilingual substitutes **can** possibly **improve** WSI results

However, **non-contextualized LM** like fasttext are **not the best choice** for this task

Appendix: semeval-2010 results

System	VM (%) (All)	VM (%) (Nouns)	VM (%) (Verbs)	#Cl
Hermit	16.2	16.7	15.6	10.78
UoY	15.7	20.6	8.5	11.54
KSU KDD	15.7	18	12.4	17.5
Duluth-WSI	9	11.4	5.7	4.15
Duluth-WSI-SVD	9	11.4	5.7	4.15
Duluth-R-110	8.6	8.6	8.5	9.71
Duluth-WSI-Co	7.9	9.2	6	2.49
KCDC-PCGD	7.8	7.3	8.4	2.9
KCDC-PC	7.5	7.7	7.3	2.92
KCDC-PC-2	7.1	7.7	6.1	2.93
Duluth-Mix-Narrow-Gap	6.9	8	5.1	2.42
KCDC-GD-2	6.9	6.1	8	2.82
KCDC-GD	6.9	5.9	8.5	2.78
Duluth-Mix-Narrow-PK2	6.8	7.8	5.5	2.68
Duluth-MIX-PK2	5.6	5.8	5.2	2.66
Duluth-R-15	5.3	5.4	5.1	4.97
Duluth-WSI-Co-Gap	4.8	5.6	3.6	1.6
Random	4.4	4.2	4.6	4
Duluth-R-13	3.6	3.5	3.7	3
Duluth-WSI-Gap	3.1	4.2	1.5	1.4
Duluth-Mix-Gap	3	2.9	3	1.61
Duluth-Mix-Uni-PK2	2.4	0.8	4.7	2.04
Duluth-R-12	2.3	2.2	2.5	2
KCDC-PT	1.9	1	3.1	1.5
Duluth-Mix-Uni-Gap	1.4	0.2	3	1.39
KCDC-GDC	7	6.2	7.8	2.83
MFS	0	0	0	1
Duluth-WSI-SVD-Gap	0	0	0.1	1.02

System	FS (%) (All)	FS (%) (Nouns)	FS (%) (Verbs)	#Cl
MFS	63.5	57.0	72.7	1
Duluth-WSI-SVD-Gap	63.3	57.0	72.4	1.02
KCDC-PT	61.8	56.4	69.7	1.5
KCDC-GD	59.2	51.6	70.0	2.78
Duluth-Mix-Gap	59.1	54.5	65.8	1.61
Duluth-Mix-Uni-Gap	58.7	57.0	61.2	1.39
KCDC-GD-2	58.2	50.4	69.3	2.82
KCDC-GDC	57.3	48.5	70.0	2.83
Duluth-Mix-Uni-PK2	56.6	57.1	55.9	2.04
KCDC-PC	55.5	50.4	62.9	2.92
KCDC-PC-2	54.7	49.7	61.7	2.93
Duluth-WSI-Gap	53.7	53.4	53.9	1.4
KCDC-PCGD	53.3	44.8	65.6	2.9
Duluth-WSI-Co-Gap	52.6	53.3	51.5	1.6
Duluth-MIX-PK2	50.4	51.7	48.3	2.66
UoY	49.8	38.2	66.6	11.54
Duluth-Mix-Narrow-Gap	49.7	47.4	51.3	2.42
Duluth-WSI-Co	49.5	50.2	48.2	2.49
Duluth-Mix-Narrow-PK2	47.8	37.1	48.2	2.68
Duluth-R-12	47.8	44.3	52.6	2
Duluth-WSI-SVD	41.1	37.1	46.7	4.15
Duluth-WSI	41.1	37.1	46.7	4.15
Duluth-R-13	38.4	36.2	41.5	3
KSU KDD	36.9	24.6	54.7	17.5
Random	31.9	30.4	34.1	4
Duluth-R-15	27.6	26.7	28.9	4.97
Hermit	26.7	24.4	30.1	10.78
Duluth-R-110	16.1	15.8	16.4	9.71