Crowd-Powered Concept Sorting for Lexicon Creation

Steve Wilson, Yiting Shen, and Rada Mihalcea, University of Michigan

{steverw,yiting,mihalcea}@umich.edu

Introduction

- Content analysis of large text corpora is a useful first step in **understanding, at a high level, what people are talking or writing about**.
- Unsupervised approaches (e.g., topic models), while useful, don't allow for much control over the **specific types of categories** being measured.
- We propose to represent lexicons using a **hierarchical tree structure** in which any node can be represented by a combination of the nodes that are its descendents. This approach:
  - Allows for **explicit modeling of hierarchical relationships**
  - Facilitates a **configurable level of specificity** when defining word categories
- However, creating and sorting the lexical hierarchy requires a great deal of manual effort, so we introduce a **crowd-powered algorithm to construct a concept tree**.
- We illustrate this process with the creation of **a lexicon to measure concepts related to personal values** [1].

Seed Terms

- We collect a set of terms that are known to be related to the target construct: personal values. We consider the following data sources:

Mobile Phone Surveys

- Asked for three values most important in people's lives
- Distributed using the mSurvey platofrm
- 1,500 total participants from: Kenya, Philippines, and Trinidad & Tobago

Online Values Surveys

- Participants wrote for 6 minutes about their values
- Extracted most common words and phrases
- Distributed via Amazon Mechanical Turk
- 1,500 total participants from: USA and India

Abridged Values Surveys

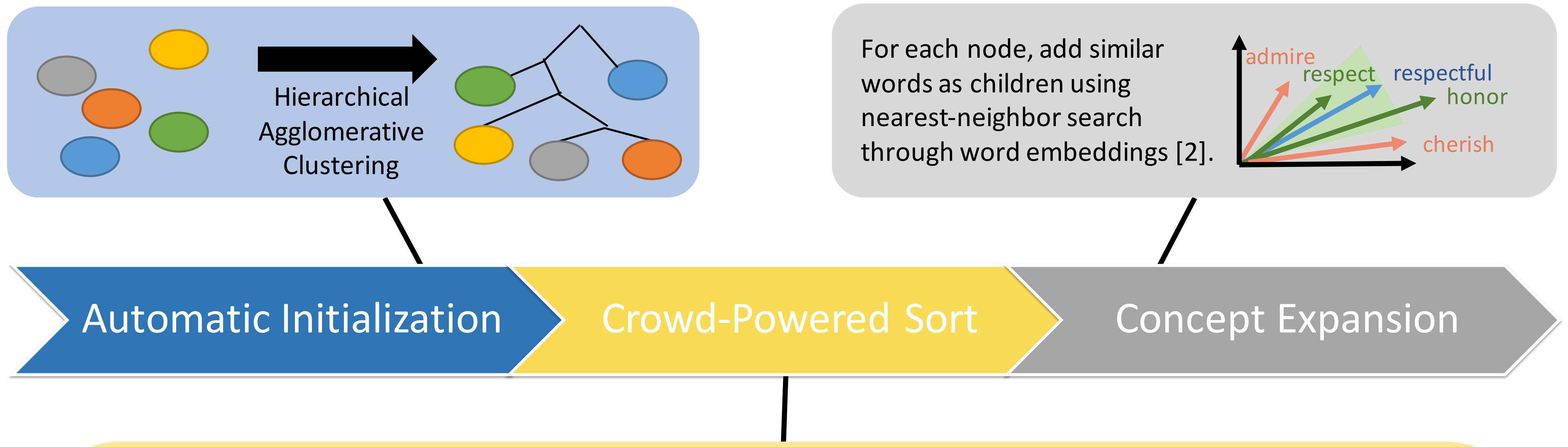
- Asked for three values most important in people's lives
- Distributed using Amazon Mechanical Turk
- 1,000 total participants from: USA and India

Templeton Values

- List of 50 common human values

Figure 1: Word cloud representing seed terms from all data sources.

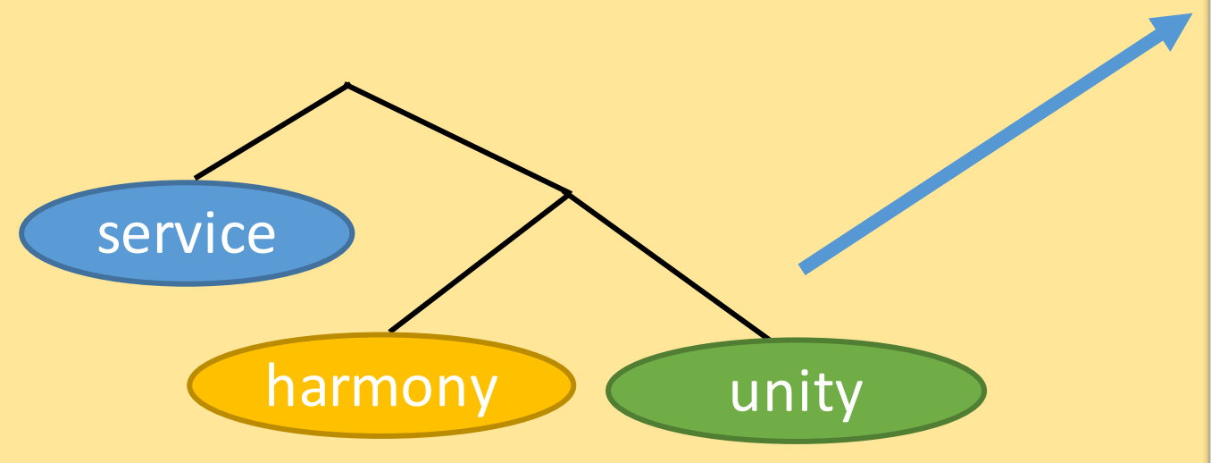
Hierarchy Construction

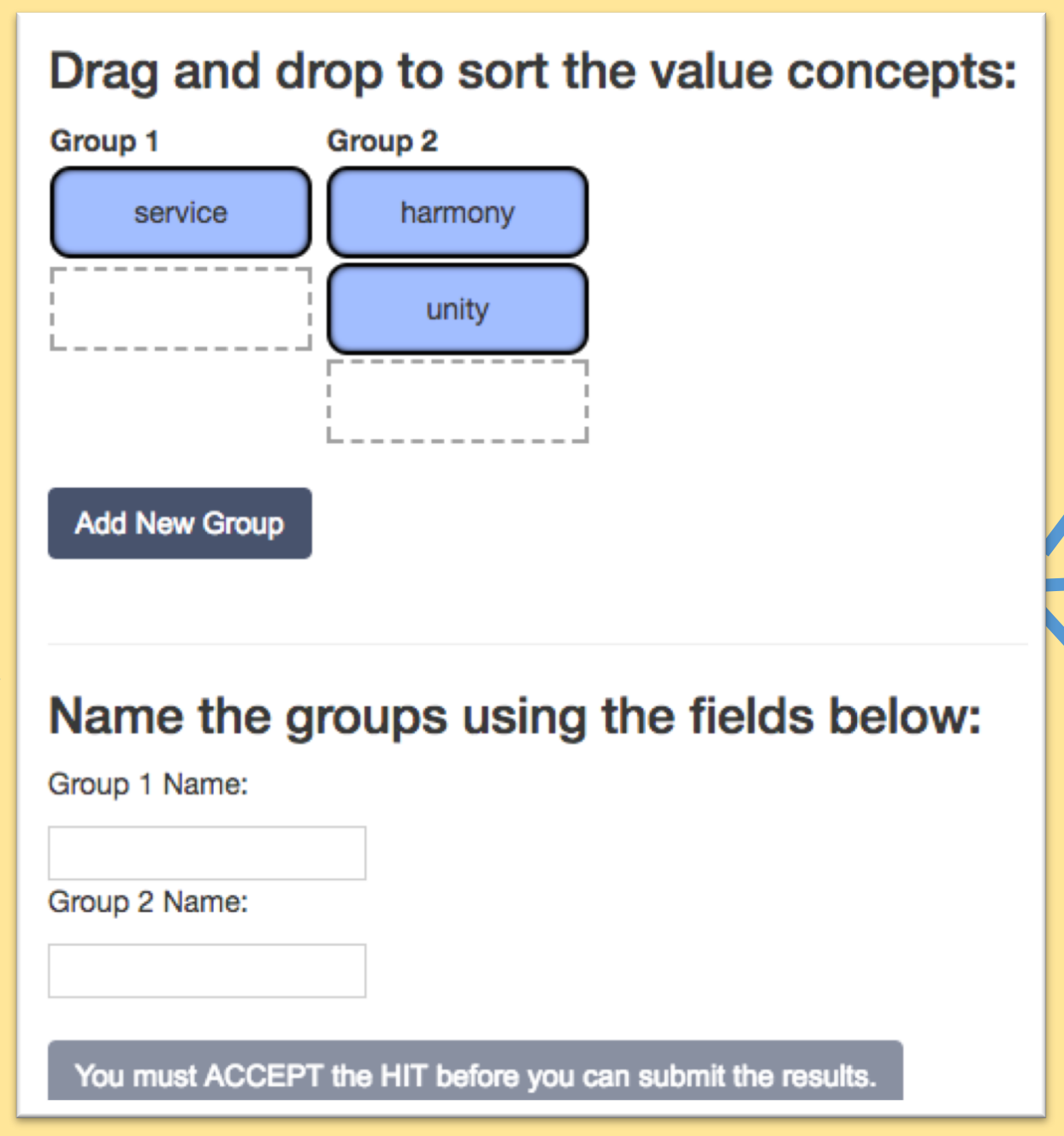


For each node, add similar words as children using nearest-neighbor search through word embeddings [2].

Automatic Initialization → Crowd-Powered Sort → Concept Expansion

- Each subtree is sorted, bottom-up, until the full tree is sorted
- Unsorted subtrees are converted into AMT sorting tasks
- Configuration chosen by the majority of workers is applied to the hierarchy.





Drag and drop to sort the value concepts:

Group 1: service

Group 2: harmony, unity

Add New Group

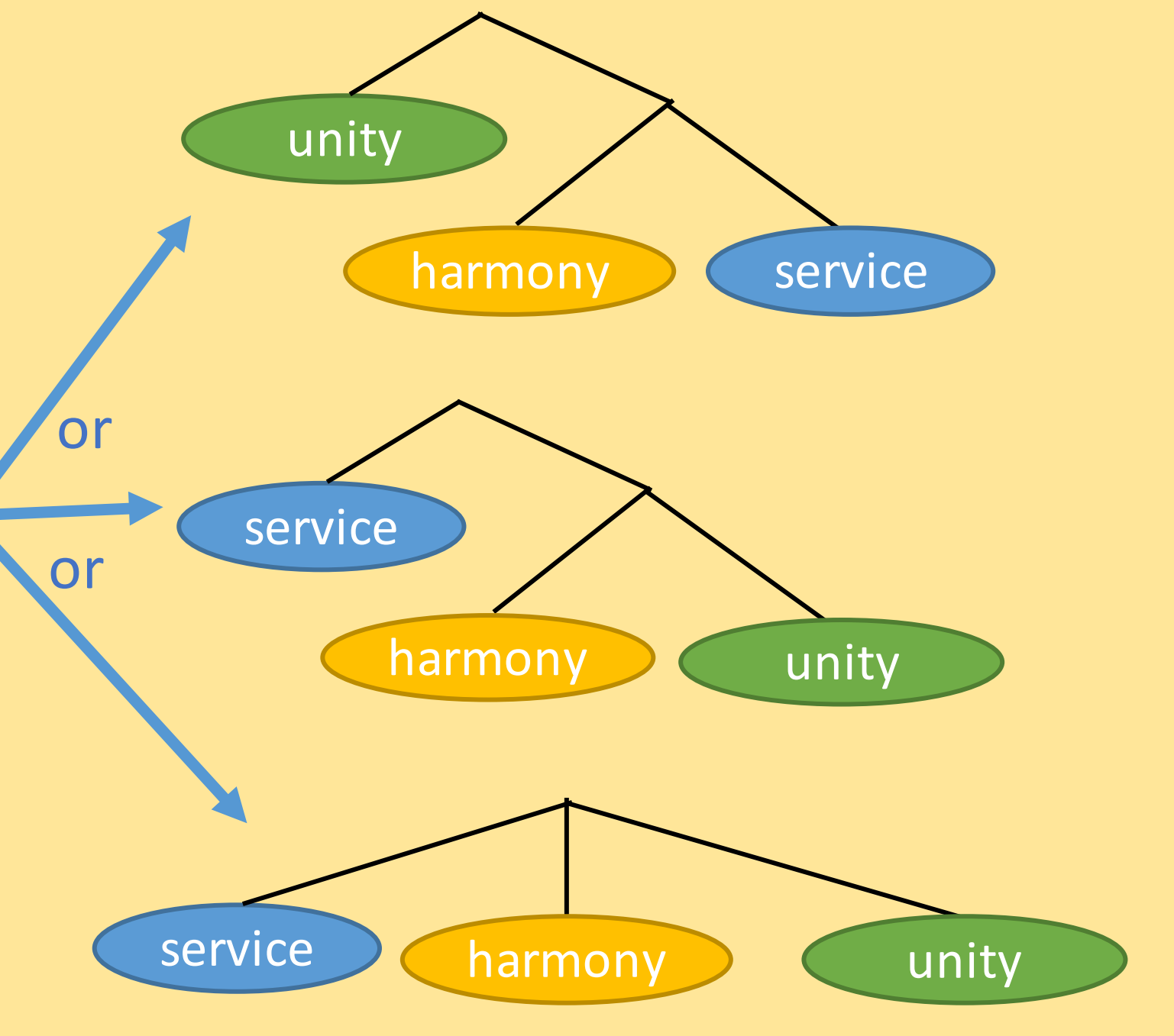
Name the groups using the fields below:

Group 1 Name:

Group 2 Name:

You must ACCEPT the HIT before you can submit the results.

Figure 2: AMT Sorting Interface.



Lexicon Evaluation

1) *Does the lexicon produce reasonable scores for documents that are known beforehand to be related to the theme of the lexicon?*

	Cognition	Emotion	Family	Learning	Optimism	Relationships	Religion	Respect	Society	Wealth
/r/christian	1.96	0.68	0.92	0.56	0.19	1.82	**6.26**	1.51	3.74	0.48
/r/college	1.34	0.57	0.39	**3.73**	0.10	0.95	0.26	1.79	3.08	1.26
/r/finance	1.29	0.29	0.09	1.26	0.17	0.58	0.04	1.01	2.07	**3.20**
/r/family	1.54	0.60	5.58	0.60	0.10	**7.20**	0.10	2.04	3.55	0.89
/r/love	2.63	1.21	0.39	0.33	0.23	1.79	0.85	1.75	**4.72**	0.39
/r/mentalhealth	2.43	1.20	0.57	0.40	0.18	1.12	0.05	1.62	**3.77**	0.73
/r/mom	1.36	0.50	4.38	0.51	0.10	**5.08**	0.08	1.73	3.93	0.91
/r/money	1.58	0.16	0.42	0.61	0.06	0.91	0.00	1.13	2.94	**5.29**
/r/parenting	1.23	0.38	3.92	0.68	0.12	**5.08**	0.10	1.78	2.76	0.81
/r/positivity	2.35	1.05	0.36	0.46	2.74	1.13	0.48	1.40	**4.71**	0.64
/r/work	1.25	0.38	0.21	0.44	0.10	0.73	0.03	1.75	**2.98**	1.22

Table 1: Frequency scores measured using selected categories

2) *Are the categories in the lexicon comprised of semantically coherent sets of words?* [3] →

3) *Do the categories in the lexicon actually measure meaningful concepts?* →

Category	MP	MPCT	CTMhl	CTMhm	**Category**	MP	MPCT	CTMhl	CTMhm
Accepting-others	0.68	1.40	0.74	0.43	Achievement	0.82	1.16	0.93	0.75
Advice	0.72	1.16	0.63	0.44	Animals	0.96	0.59	0.86	0.93
Art	1.00	0.92	0.83	0.50	Autonomy	0.80	0.80	0.50	0.83
Career	0.90	1.13	1.00	0.96	Children	0.94	1.14	0.91	1.00
Cognition	0.94	1.32	0.76	0.44	Creativity	0.84	1.02	0.64	0.73
Dedication	0.92	1.39	0.85	0.50	Emotion	0.82	1.29	0.68	0.46
Family	0.95	0.87	0.85	1.00	Feeling-good	0.92	1.01	0.70	0.69
Forgiving	0.90	1.02	0.64	0.95	Friends	0.74	0.92	0.65	0.72
Future	0.62	1.29	0.58	0.65	Gratitude	0.94	0.93	0.42	0.64
Hard-work	0.90	1.01	0.71	0.52	Health	0.96	0.43	0.71	0.95
Helping-others	0.86	1.37	0.36	0.31	Honesty	0.94	1.07	0.67	0.78
Inner-peace	0.70	1.01	0.96	0.24	Justice	0.82	1.29	0.43	0.39
Learning	0.84	0.86	0.97	0.61	Life	0.74	1.27	0.89	0.26
Marriage	0.80	0.90	0.93	0.69	Moral	0.92	1.19	0.54	0.67
Optimism	0.84	0.93	0.96	0.91	Order	0.90	1.05	0.54	0.30
Parents	0.80	0.99	0.77	0.91	Perseverance	0.94	1.04	0.68	0.23
Purpose	0.64	0.83	0.38	0.30	Relationships	0.92	1.06	1.00	0.78
Religion	0.66	1.26	1.00	1.00	Respect	0.36	1.03	0.11	0.48
Responsible	0.60	1.06	0.77	0.65	Security	0.78	1.11	0.83	0.64
Self-confidence	0.78	0.91	0.85	0.75	Siblings	0.68	0.91	1.00	1.00
Significant-others	0.89	0.81	0.71	0.73	Social	0.63	1.11	0.84	0.75
Society	0.68	0.69	0.07	0.54	Spirituality	0.68	0.85	0.65	0.83
Thinking	0.90	1.37	1.00	0.92	Truth	0.68	1.11	0.63	0.81
Wealth	0.96	0.69	1.00	0.92	Work-ethic	0.86	1.15	0.45	0.50
					Baseline	*0.33*	*0.00*	*0.50*	*0.50*
					Average	**0.81**	**1.04**	**0.66**	**0.72**

Table 2: Model Precision, Model Precision Choose-Two, and Category-Text Matching Scores for all categories.

Sample Categories

- **LEARNING:** profs colleges educate educators researches faculty schooling professors scholastic college learning lesson schoolhouse campus lessons educational…
- **WORK-ETHIC:** duty perseverance motivation tough hardworking chore endeavor accountability perseverence industrious strength work_hard…
- **HELPING-OTHERS:** supporting help_the_needy aiding make_a_difference another aids further do_no_harm succour giving support contributed contributing other…
- **AUTONOMY:** independently independent autonomy sovereign independant independents self-motivation self-sufficiency self-reliance freelance automated…
- **ACHIEVEMENT:** achievements successful productivity succeeded success attainment successes conquest accomplishment avail efficiency accomplishments…

Sorting Algorithm Details

```
Algorithm 1: Crowd-powered Tree Sorting.
Data: $\mathcal{T}$: Tree to be sorted, $n$: number of annotators, $m$: maximum HIT
      extensions
Result: $\mathcal{T}'$: Sorted Tree
Function traverseAndSortTree($\mathcal{T}$, $n$, $m$)
    if numChildren ($\mathcal{T}$) $> 0$ then
        foreach $\mathcal{S} \in$ DirectSubtrees ($\mathcal{T}$) do
            $S \leftarrow$ traverseAndSortTree($\mathcal{S}$, $n$, $m$));
        $\mathcal{T}' \leftarrow$ sortSubtree ($\mathcal{T}$, $n$, $m$);
        foreach $\mathcal{U} \in$ (DirectSubtrees ($\mathcal{T}'$) \ DirectSubtrees ($\mathcal{T}$)) do
            $\mathcal{U} \leftarrow$ traverseAndSortTree ($\mathcal{U}$, $n$, $m$);
    else
        $\mathcal{T}' \leftarrow \mathcal{T}$;
    return $\mathcal{T}'$;
Function sortSubtree ($\mathcal{T}$, $n$, $m$)
    $G \leftarrow$ makeGroups (DirectSubtrees ($\mathcal{T}$));
    $H \leftarrow$ createHIT ($G$);
    $n' \leftarrow n$;
    $s \leftarrow 0$;
    while $!s$ do
        $R \leftarrow$ checkHITResults ($H$);
        if $|R| \geq n'$ then
            if majorityAgree ($R$) or $n' \geq (m+1) \times n$ then
                $s \leftarrow 1$;
                $\mathcal{T}' \leftarrow$ mostCommon ($R$);
            else
                $H \leftarrow$ extendHIT ($H$, $n$);
                $n' \leftarrow n' + n$;
    return $\mathcal{T}'$;
$\mathcal{T}' \leftarrow$ traverseAndSortTree($\mathcal{T}$, $n$, $m$);
```

References

[1] Wilson, S., Shen, Y., and Mihalcea, R. Building and Validating Hierarchical Lexicons with a Case Study on Personal Values. To appear in *proceedings of the 10th International Conference on Social Informatics*. (2018)

[2] Mrkšić, N., Séaghdha, D.O., Thomson, B., Gašić, M., Rojas-Barahona, L., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Counter-fitting word vectors to linguistic constraints. arXiv preprint arXiv:1603.00892 (2016)

[3] Morstatter, F., Liu, H.: A novel measure for coherence in statistical topic models. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers). vol. 2, pp. 543–548 (2016)

Acknowledgements

This material is based in part upon work supported by the National Science Foundation (#1344257), the John Templeton Foundation (#48503), the Michigan Institute for Data Science, and the National Institute of Health (#5R01GM112697-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, the Michigan Institute for Data Science, or the National Institute of Health.