Measuring Semantic Relations between Human Activities

Li Zhang, Steve Wilson, and Rada Mihalcea, University of Michigan
{zharry,steverw,mihalcea}@umich.edu

Introduction

- Our everyday activities say a lot about who we are:
  - Personality [1]
  - Interests [3]
  - Values [2]
  - Future actions [4]
- We can't always directly observe human activities, yet people talk about what they are doing online. Examples:
  - Tweets
  - Facebook status updates

- However, reasoning about the relationships between activity phrases is not always straightforward:
  - *go to a bar / attend church* (noun is important)
  - *exercise / hit a punching bag* (type-of relationship)
  - *sell a car / drive an SUV* (verb is important)
  - *drink coffee / eat breakfast* (often done together)
- Our goal: Build a model that is able to determine the semantic relationship between pairs of activity phrases:

Data

- To evaluate how well computational models are able to capture relationships between human activities, we create the Human Activity Dataset [5].
- Pairs of activities were annotated across four dimensions:

Similarity
- Semantic similarity in a strict sense.
- Example of high similarity phrases: *to watch a film* and *to see a movie*.

Relatedness
- A general semantic association between two phrases.
- Example of strongly related phrases: *give a gift* and *receive a present*.

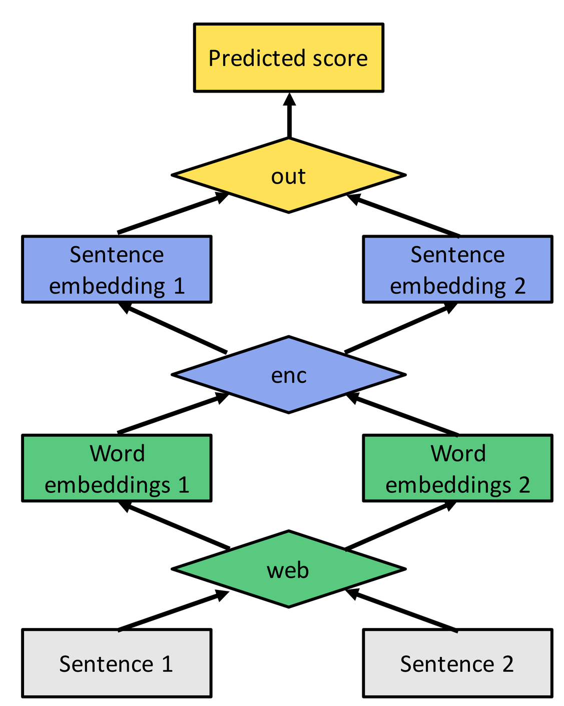
Motivational Alignment
- The degree to which the activities are done with similar motivations.
- Example of phrases with potentially similar motivations*: eat dinner with family members* and *visit relatives*.

Perceived Actor Congruence
- Is someone who often does an activity also expected to do a second activity?
- Example of activities with high PAC score: *travel* and *pack a suitcase*.

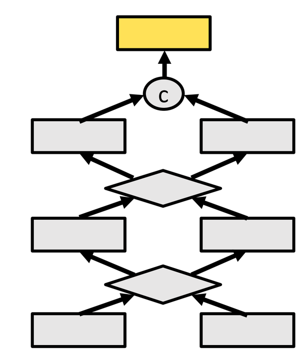
Activity 1	Activity 2	SIM	REL	MA	PAC
go jogging	lift weights	1.67	2.22	2.89	1.11
read to one's kids	go to a bar	0	0	0	-1.29
take transit to work	commute to work	3.38	3.5	3.38	0.5
make one's bed	organize one's desk	0.58	1.29	1.57	0.71

Table 1: Sample of scores assigned to pairs of activities in the Human Activity Dataset. SIM, REL, and MA scores are in the range [0,4] and PAC scores lie in [-2,2]. Scores are averaged across 10 annotators.

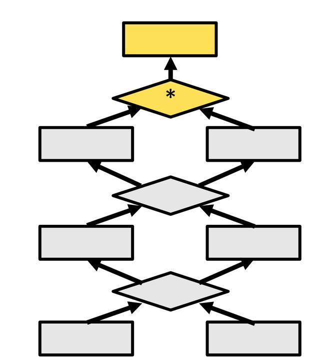
Sentence Encoder Transfer Settings



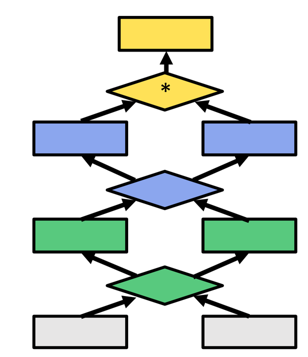
General Sentence Encoder Training Architecture
- Initial parameters are learned from a large *source* dataset such as SNLI [6] or Back-translated text [7].
- Afterwards, there are several ways to transfer the learned model to the *target* dataset (i.e., human activities).



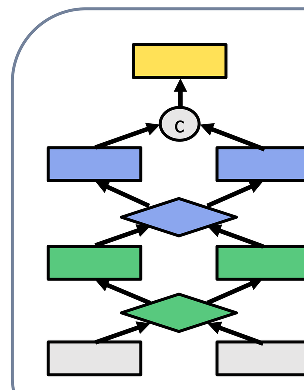
Unsupervised Evaluation
- No parameters are updated using the target dataset
- Labels are generated directly from sentence embeddings using c (cosine similarity)



Feature Transfer
- All embedding layers are frozen during target training
- A new logistic regression classifier * is trained to make predictions



Network Transfer
- Embedding parameters are allowed to be updated during target training
- A new logistic regression classifier * is trained to make predictions



Direct Network Transfer
- Embedding parameters are allowed to be updated during target training
- Labels are generated **directly** from sentence embeddings using c (cosine similarity)

Experimental Results and Analysis

Transfer Experiments

- We use the pre-trained InferSent model [8] to initialize our model parameters before transferring to each of the four dimensions in the Human Activity Dataset.

Transfer Setting	SIM	REL	MA	PAC
Unsupervised Evaluation	**.701**	.686	.652	.525
Feature Transfer	.655	.644	.608	.432
Network Transfer	**.699**	.692	.672	.537
Direct Network Transfer	**.702**	**.722**	**.691**	**.572**
Human Agreement	*.768*	*.768*	*.745*	*.620*

Table 2: Spearman's correlation between model predictions and human judgments for the four relational dimensions

- Direct Network Transfer is especially helpful when transferring to less traditional relational dimensions such as MA and PAC.

When Transfer Works

- We distinguish between two types of pairs for which transfer helps and show some illustrative examples:

1. Pairs with scores that were initially overestimated
2. Pairs with scores that were initially underestimated

	Phrase 1	Phrase 2
1.	have dinner with friends	eat dinner by oneself
1.	go to a party	go to bible study
1.	play football	play basketball
1.	go to the movie theater	go to office to work
2.	take long walks	go on a walk
2.	take care of one's dogs	groom one's dog
2.	read books	visit a bookstore
2.	go to the doctor	see the doctor

Importance Analysis

- We use the leave-one-out importance analysis introduced in [9] as a basis for the following definition of the irrelevance of a word w for model m_1 trained only on the source data and model m_2 after transferring to the target data:

$$irrelevance(w, p_1, p_2, m_1, m_2) = m_2(p_1^{\neg w}, p_2) - m_1(p_1^{\neg w}, p_2)$$

where p_1 and p_2 are phrases that form a training instance, $p^{\neg w}$ is phrase p with the word w removed, and $m(p_1, p_2)$ is the model's prediction of the relationship between p_1 and p_2.

- This allows us to quantify the extent to which the model treats each word different after transfer.
- Using this approach, we explore the effect of Direct Network Transfer to the PAC dimension.

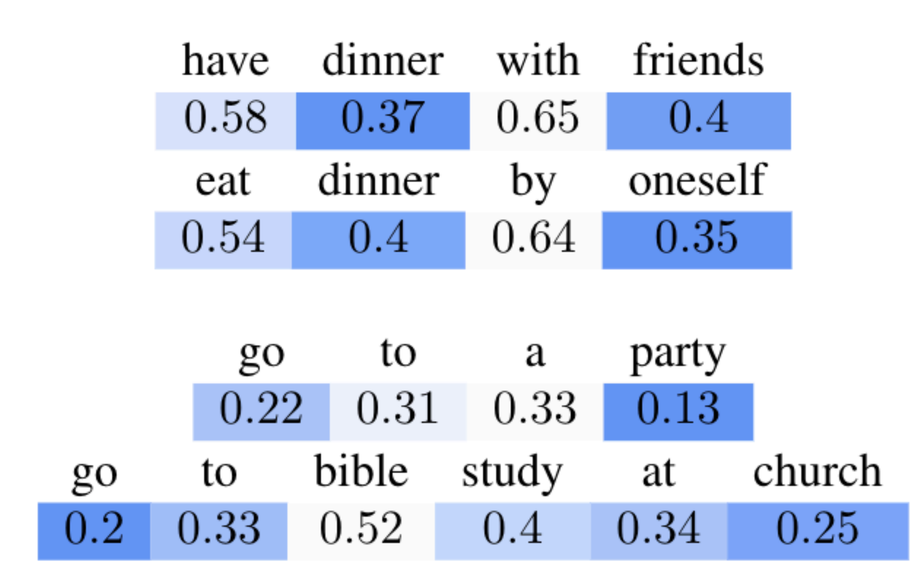


Figure 1: Heatmap of irrelevance scores showing the effect of Direct Network Transfer to the PAC dimension. Darker boxes indicate words that became more relevant during transfer.

Transfer to the STS Benchmark

- We also test the ability of Direct Network Transfer to fine-tune models for other datasets, such as the Semantic Text Similarity Benchmark [9]:

Transfer Setting	Dev	Test
Unsupervised Evaluation	.791	.783
Feature Transfer	.779	.746
Network Transfer	.836	.810
Direct Network Transfer	**.852**	**.824**
Previous Best [10]	*.847*	*.810*

Table 3: Pearson correlation with ground truth labels on the STS Benchmark evaluation.

Conclusions

- Our Human Activity Dataset [5] serves as a **resource for the training and evaluation** of semantic similarity methods in the domain of **human activities**.
- Experimental results show that **transfer learning** allows us to accurately model the relationships between human activities by **leveraging information learned from from very large text corpora**, even if the domain varies.
- We introduce the ***Direct Network Transfer*** setting, which gives the best results on the Human Activity Dataset and is **successful on other datasets**, including state-of-the-art performance on the STS Benchmark.

References

[1] Icek Ajzen. 1987. Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. *Advances in experimental social psychology* 20:1–63.
[2] Milton Rokeach. 1973. The nature of human values. Free press.
[3] Jeremy Goecks and Jude Shavlik. 2000. Learning users' interests by unobtrusively observing their normal behavior. *In Proceedings of the 5th international conference on Intelligent user interfaces*. ACM, pages 129–132.
[4] Judith A Ouellette and Wendy Wood. 1998. Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological bulletin* 124(1):54.
[5] Wilson, S. and Mihalcea, R., 2017. Measuring Semantic Relations between Human Activities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Vol. 1, pp. 664-673).
[6] Bowman, S.R., Angeli, G., Potts, C. and Manning, C.D., 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
[7] Wieting, J., Mallinson, J. and Gimpel, K., 2017. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847*.
[8] Conneau, A., Kiela, D., Schwenk, H., Barrault, L. and Bordes, A., 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.
[9] http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark
[10] Tian, J., Zhou, Z., Lan, M. and Wu, Y., 2017. ECNU at SemEval-2017 Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-lingual Semantic Textual Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 191-197).

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.