Luminoso at SemEval-2018 Task 10:

Distinguishing attributes using text corpora and relational knowledge

Robyn Speer and **Joanna Lowry-Duda**

Abstract

Luminoso participated in the SemEval 2018 task on “Capturing Discriminative Attributes” with a system based on ConceptNet, an open knowledge graph focused on general knowledge. We describe how we trained a linear classifier on a small number of semantically-informed features to achieve an F_1 score of 0.7368 on the task, achieving second place on the post-evaluation leaderboard.

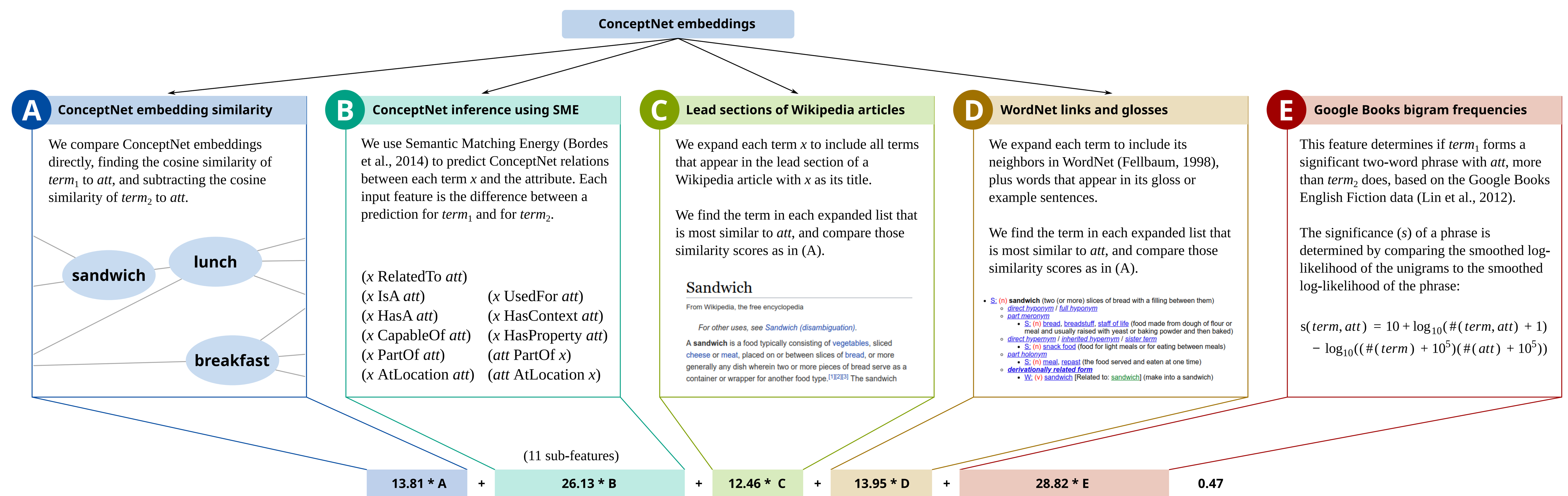
Task description

The task is to identify attributes that are typically associated with the first of a pair of words and not the second.

Examples

Term 1	Term 2	Attribute	Discriminative?
lambs	cattle	wool	✔
Lambs produce wool, while cattle do not.			
shoulder	leg	arm	✔
A shoulder is attached to an arm, while a leg is not.			
train	subway	rails	✖
Both a train and a subway involve rails, so rails are not a discriminative attribute here.			
finger	soup	water	✖
Soup may be related to water, but this is the wrong direction. In this task, a discriminative attribute must be related to the first term and not the second.			

What forms of input help to distinguish attributes?



ConceptNet embeddings

This is a task based on general knowledge, with a small amount of training data. Solving the task requires a model of general knowledge that cannot be learned at training time.

For this, we used ConceptNet embeddings, similar to those that won SemEval 2017 task 2 (Speer and Lowry-Duda, 2017).

These embeddings are used directly as feature *A*, used as the initial input layer of the externally-trained semantic model *B*, and used for semantic comparisons in *C* and *D*.

Avoiding overfitting

To minimize the number of free parameters and therefore the potential for overfitting to the small training set, we trained a simple linear SVM model, on 15 input features from 5 sources.

We took advantage of the design of our features and the asymmetry of the task as a way to further mitigate overfitting. All of the features were designed to identify an attribute that $term_1$ has and $term_2$ does not. Any feature with a negative weight, therefore, purely represents overfitting on the training data. Setting negative weights to 0 after training yields a more robust classifier.

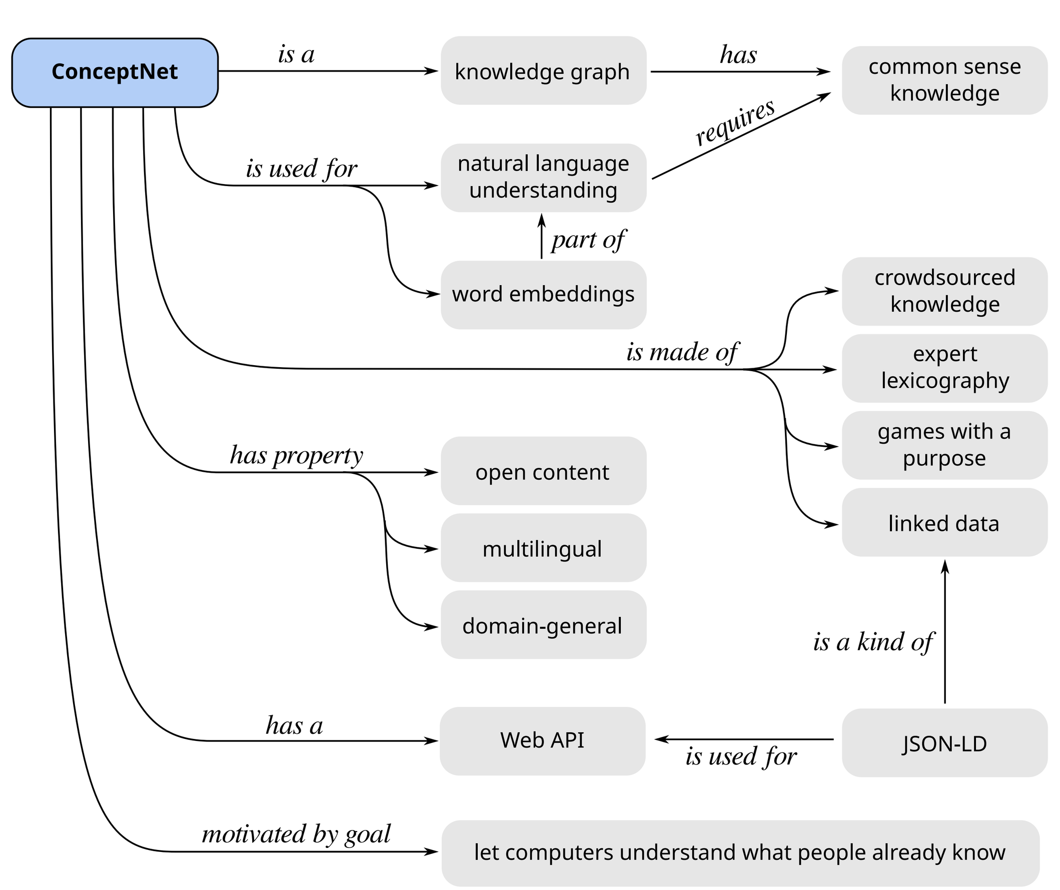
Classifier parameters

We used LinearSVC, an implementation of liblinear (Fan et al., 2008) within scikit-learn (Pedregosa et al., 2011).

The SVC parameters were the defaults for scikit-learn 0.19:

- Soft margin: $C = 1.0$
- Squared hinge loss
- L_2 penalty on coefficients
- Solving the dual form of SVM

What is ConceptNet?



Data sources in ConceptNet

We used the embeddings generated by ConceptNet 5.5.5 in their entirety for this task. It can be useful to know where ConceptNet's input data came from:

Crowdsourcing

- Open Mind Common Sense
- OMCS no Brasil
- Wiktionary
- Wikipedia via DBPedia

Expert resources

- Open Multilingual WordNet
- JMDict
- CEDict
- OpenCyc
- Unicode CLDR emoji data

Games with a Purpose

- Verbosity (English)
- nadya.jp (Japanese)
- PTT Pet Game (Chinese)

Distributional semantics

- word2vec, precomputed on Google News
- GloVe, precomputed on the Common Crawl
- fastText, customized to learn from parallel text, trained on OpenSubtitles 2016

The details of how ConceptNet is built, and individual citations for its data sources, appear in the AAAI paper on ConceptNet 5.5 (Speer et al., 2017).

ConceptNet is all you need

Our full classifier used the linear combination of 5 types of input features shown above. This point is labeled **ABCDE** on the graph to the right. The other points are ablated versions of the classifier, trained on subsets of the five sources.

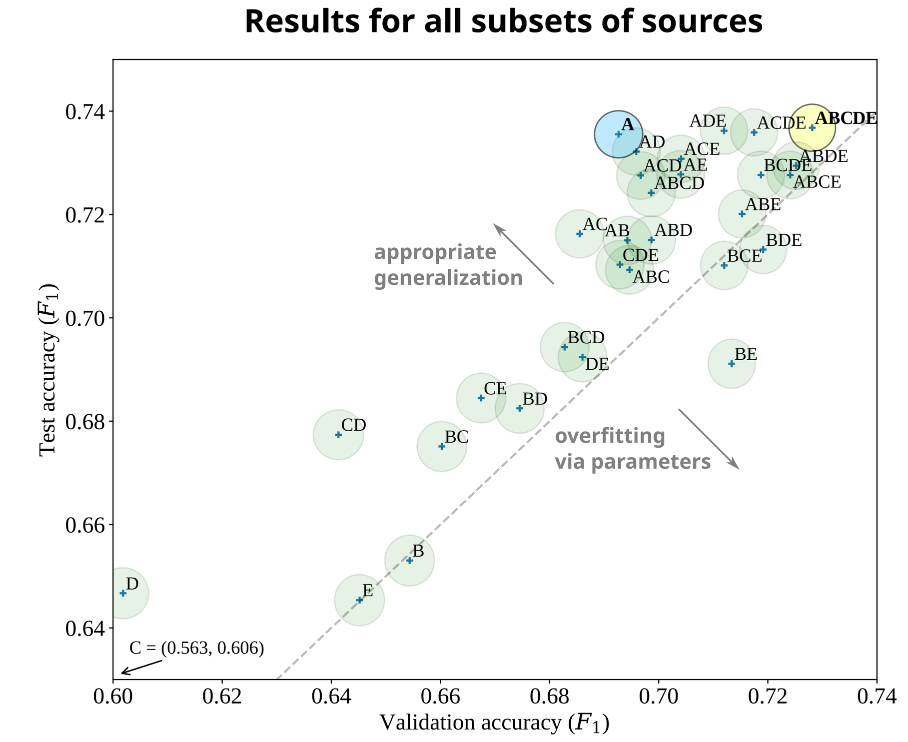
We found that the single feature of ConceptNet similarity **(A)** performed just as well on the test data as the full classifier, despite its lower validation accuracy.

This one-feature classifier could be more simply described as a heuristic over cosine similarities of ConceptNet embeddings:

$$\text{sim}(term_1, att) - \text{sim}(term_2, att) > 0.0961$$

It seems that the test data contained distinctions that can already be found by comparing ConceptNet embeddings, and that more complex features may have simply provided an opportunity to overfit to the validation set by parameter selection.

Results for all subsets of sources



This graph shows the validation and test accuracy of classifiers trained on subsets of the five sources of features. Ellipses indicate standard error of the mean, assuming that the data is sampled from a larger set.

References

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2014. A semantic matching energy function for learning with multi-relational data. Machine Learning, 94(2):233–259.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classif cation. Journal of Machine Learning Research, 9(Aug):1871–1874.

Christiane Fellbaum (ed.) WordNet: A Lexical Database for English. MIT Press Cambridge.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. In Proceedings of the ACL 2012 system demonstrations, pages 169–174. Association for Computational Linguistics.

Denis Paperno, Alessandro Lenci, and Alicia Krebs. 2018. SemEval-2018 Task 10: Capturing discriminative attributes. In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, United States. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct):2825–2830.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In AAAI, San Francisco.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 85–89, Vancouver, Canada. Association for Computational Linguistics.

Wikipedia. 2017. Wikipedia, the free encyclopedia — English data export. (A collaborative project with thousands of authors.) Retrieved from https://dumps.wikimedia.org/enwiki/ on 2017-12-20.

Open code and data

Code: https://github.com/LuminosoInsight/semeval-discriminatt
Data: http://zenodo.org/record/1183358

The Zenodo link contains an archive of all of our input data. Together with the code repository on GitHub, it enables reproducing the result presented here.

ConceptNet can be browsed and downloaded from http://conceptnet.io.