ParaLex: A Multilingual Resource for Evaluating Semantic Similarity Models

Peter Sumbler, Fredrik Olsson, Nina Viereckel, Maria Verbitskaya, Lars Hamberg, Magnus Sahlgren, and Jussi Karlgren

Gavagai
www.gavagai.se
Stockholm

Abstract. This paper introduces a novel dataset and methodology for evaluating semantic similarity models in different languages. The dataset, called ParaLex, consists of thematic word lists that represent a number of different semantic paradigms, such as days of the week, months of the year and common fruit and vegetables, in 46 different languages. We describe the design choices made when creating these paradigms, and the problems we have encountered translating and verifying them in a large number of different languages. We release our dataset in a GitHub repository (`https://github.com/gavagai-corp/ParaLex/`) and invite public contribution and collaboration. We also introduce an evaluation methodology that utilizes these word lists for measuring the suitability of a semantic similarity model for populating semantically coherent term sets. We offer a comparison of our evaluation with other notable evaluation benchmarks in English. We then exemplify the evaluation procedure with a number of established multilingual semantic similarity models: embeddings from fastText, Polyglot and ConceptNet.

Keywords: Semantic Similarity · Evaluation · Lexical Resources · Word Embeddings · Multilingual Resources.

1 Evaluating Semantic Models

The use of distributional semantic models implemented as word embeddings is now near ubiquitous and successful in a majority of language processing tasks. There is an attendant and growing interest in the evaluation of the quality of such semantic models, and some intrinsic evaluation benchmarks have become a de-facto standard, including: similarity tests, such as SimLex-999 [16], SimVerb-3000 [12] and WordSim-353 [11]; multiple-choice synonym tests, such as the TOEFL synonym test set [19] and ESL [27]; more in-depth evaluation of the representation of specific semantic relations such as BLESS [3] and EVALution [24]; and testing of the encoding of syntactic and semantic analogies into the vector space [21, 15]. The similarity and synonym measures are primarily designed to emulate human semantic competence and evaluate semantic proficiency. Analogy tests are primarily motivated by structural concerns, and are often used to gauge a model's suitability as a representation layer in machine learning.

Industrial practice, on the other hand, may pose other requirements on the evaluation, and often requires different evaluation schemes than academic research experimentation. The proposed evaluation is based on three concerns.

Firstly, we require evaluation for a large number of languages, 46 at time of writing. Although developing test data sets in specific individual languages is an ongoing effort in the research community [9], these lack standardization, uniformity and consistency on a multilingual level. While multilingual evaluation projects indeed exist, we have not yet found such a project with sufficient languages for our particular case.

Secondly, the correlation of intrinsic metrics with success in downstream language-based tasks is unclear, and even for those that do correlate, it is with target tasks which are ill defined or of dubious commercial utility [25, 2, 4, 7, 13, 20, 6, 23]. Faruqui et al. (2016) even go as far as to suggest that the evaluation of distributional models using word similarity tasks is "not sustainable" in the future, both as a combination of correlation, annotation and task-design flaws. We note that current conclusions around intrinsic evaluation vs. downstream tasks are messy and follow ideas floated by Gladkova and Drozd (2016) and Batchkarov et al. (2016) that the field needs to move towards analyzing specific strengths and weaknesses of word embeddings, rather than distilling quality into one single metric. We design our evaluation methodology to be specific and pertinent to the style of language processing tasks we undertake in our services. We incorporate distributional information into the analysis of large amounts of unstructured textual data and need an evaluation measure which assesses how well our implementation identifies closely related concepts. To our knowledge, no equivalent resource exists thus far.

Thirdly, an evaluation procedure should be as interpretable and understandable as possible, leading the system development process towards useful enhancements and fixes. What is more, given that machines are not currently so intelligent that they can carry out extensive data analysis unaided, we design our evaluation from the point of view of keeping a human analyst in the loop.

2 Evaluation Methodology

Our methodology of evaluation has been designed to mimic the way a user would go about interactively and iteratively creating a list of thematically related terms based on suggestions provided by a (distributional) semantic model. From a linguistic point of view, such a term list can be considered to be a set of co-hyponyms that can be grouped under a tangible category label, such as: days of the week, French singers or former Swedish prime ministers (not all of these feature in the actual dataset). The term "paradigm" is also an appropriate description.

In building a thematic term list, a user is invited to enter one or more initial seed terms and is then presented with suggestions, provided by the semantic model at hand, which may be suitable for inclusion in the list. If these suggestions are sensible and fit in with the paradigm, we have good reason to believe that

the model is of good quality. Our test will therefore revolve around the quality of such suggestions. We are aiming to measure how well a (distributional) semantic model captures these paradigmatic notions of similarity.

2.1 Test Architecture

The proposed test works with clusters of terms grouped together in paradigms, as detailed above. In section 3, we lay out our ParaLex dataset, which is composed of a number of such clusters and currently available in 46 different languages.

The fundamental architecture of the test involves splitting the test data paradigms into a set of seeds and a set of targets. On sampling two seed words from a cluster, the remaining items become the set of targets we would ideally like to find in the suggested terms. Starting from the initial seed or seeds, we check the current suggestions given and are able to calculate scoring metrics with respect to the target set. According to certain criteria, we accept some of these suggestions into our list of terms, which join the existing members to become a larger set of seeds for the next iteration. In each iteration we repeat the process: get suggestions, calculate the score metrics and filter suggestions to be added as seeds for the next step.

This aims to mimic the process a user goes thorough whilst constructing a list of thematically related terms, typing in some keywords, getting suggestions, then iteratively choosing which suggestions to accept into the term list and again obtaining suggestions. The filtering criteria can be considered a proxy for user behavior.

2.2 Filtering Criteria

Now we define the mechanism by which suggestions are decided to be sufficiently relevant. We accept suggestions which are in some way coherent with the current set of seeds.

We define n-coherence of a word vector with a set of seeds to mean "occurring in the local neighborhood of at least n terms". There is no theoretical method to determine the appropriate coherence threshold a-priori. In preliminary experiments, we conducted tests using varying levels of coherence.

It is important to note that a coherence threshold of one is the same as accepting all suggestions as relevant. With these parameters, the test does not operate in a way that is helpful for evaluating the quality of a semantic model. Adding every single suggestion is not representative of how a user would use the term suggester to create a term list. In addition, the amount of suggestions quickly becomes very large and the level of "common sense" coherence of the list is minimal and the list is not in any way representative of a paradigm.

Conversely, we do not want to raise the threshold too high. This would carry the danger of filtering out good suggestions, making potentially good semantic model appear worse than they are. In preliminary tests, the aim was to keep the threshold as low as possible without resorting to the messy situation of adding all

suggestions. A coherence threshold of two proved to be a reasonable compromise, being only one step stricter than the chaotic threshold of one.

2.3 Scoring

Given that we are interested in finding items with respect to a particular class (items in the target set), the natural scoring functions lending themselves to this task are precision and recall. We are more interested in recall as a score in this context, as the interest lies in how many targets we are able to find. The precision will always be low as we are returning a large number of suggestions compared to the number of targets we are looking for. Additionally, when compiling a thematic list, a user is likely to skim the list of items and pick out the relevant suggestions. In this context a number of less-than-perfect items is arguably excusable up to a certain point. This also relates to the design feature of keeping a human user in the loop.

On the other hand, if the user is presented with too many suggestions, this is also a problem. For this reason, the number of suggestions is capped at two hundred. In the case that the set of suggestions grows larger than this, the test will stop iterating and the previous current score will taken as the final one.

2.4 Further Notes

For simplicity, we will carry out 3 iterations for every test, reporting the final score.

We want to have a deterministic test which gives the same result every time, but we do not want the test to be vulnerable to the choice of seeds. We therefore permute every possible combination of seeds and targets and carry out the test for each permutation. The cluster score is the mean of every possible permutation. The overall test score is in turn the mean of all the cluster scores.

Before performing any cluster test, we must first check that each cluster term is in the vocabulary of the model under evaluation. Any missing terms have to be removed, as we cannot check suggestions for terms which do not even exist in the semantic model. In addition, we must also check for the case where there are too many out of vocabulary terms to carry out a test. There must be at least 2 seed terms and 1 target terms. Any cluster with less than 3 terms in the model vocabulary receives a zero score, as no test can take place.

The test in its current form is prescriptive by nature. The requirements that the model must meet to get a good score are set out a-priori before the test takes place. An automatic test must be defined in this way, otherwise it cannot be performed without human intervention. On the other hand, this does not necessarily tell us the whole story.

There may be a particular reason why a semantic model is not getting very good results on a given cluster. It is perfectly possible that many of the suggestions are of high quality, but do not exactly match the target suggestions as detailed in the test data. For these reasons it is essential that the paradigmatic

clusters used for evaluation are sufficiently high quality - that is, representative and complete.

3 Evaluation Data

A key component of this evaluation resource is the test data. Motivated by a desire to test how well semantic models perform at our word-list task, we were unable to find sufficiently complete test data for this purpose, particularly from a multilingual point of view. To this end, we have created ParaLex: a dataset of semantic clusters in 46 languages. Due to space concerns, the dataset is hosted in a GitHub repository.[1] Although the dataset was created to facilitate our suggested evaluation methodology, we also hope the data will be useful for other purposes, such as neighborhood coherence tests [18].

ParaLex consists of manually constructed paradigmatic groups of terms. For each language, there exist 22 or 13 distinct thematic clusters covering such topics as colors, cities, and fruits. These are intended to represent cross-linguistic conceptual categories.

Designing and creating the dataset was a non-trivial endeavor. A number of decisions were made and difficulties negotiated in the process. One particular design feature was the choice of the thematic categories to be included in the dataset. It must be ensured that the categories chosen are relevant, representative and sufficiently prevalent concepts to be taken as a marker of a general word space quality.

The clusters must also be chosen to ensure that the concepts represented can be easily rendered in each individual language, without significant translation difficulties. Cultural considerations come into play when transferring concepts to a new language. For instance, if we have a cluster of common fruits, the choice of these may be very different depending on if the language in question is Swedish or Swahili, as the commonly found fruits in these cultures where these languages are spoken do not necessarily coincide. It makes little sense to force a translation of a particular fruit if it is not a relevant member of the class in the target language. It is much more realistic to choose a different member with an equivalent level of prominence. The dataset has been built using this general principle.

This last point also applies to the choice of the clusters themselves. For example, if the language never uses abbreviations to refer to months, this cluster must be omitted. Indeed, some Asian languages simply use cardinal numbers rather than named references (e.g. month 1, month 2, ...). In such a case, it would not be relevant to carry the cluster over into the new languages. This is why there is a variable number of clusters between languages.

We can also divide our paradigms into open and closed class categories. This has been an important distinction to make. For closed classes, such as months, we have created exhaustive lists of the concept (where this is practical). For

[1] `https://github.com/gavagai-corp/ParaLex/`

reasonably open class items we have aimed for a hopefully representative sample of the concept.

The dataset in its current form focuses on concrete nouns, as this has been the most pertinent use case up until now. There is no reason why this should continue moving forward. In fact, for a dataset to be truly representative, it should ideally be balanced across all parts of speech.

In addition, given the large number of languages that our project is interested in, it is difficult to quality control each localization. The number of languages spoken by one individual seldom reaches 46. It is essential that precise instructions are given to a person carrying out cluster creation in a particular language. However, it is much easier to give clear instructions about the creation of paradigms compared to, for example, word-pair similarity datasets. Avraham and Goldberg (2016) give a good indication about just how many complications there are with the latter.

We consider our dataset to be a proof-of-concept and a work-in-progress, ripe for future development. This is one of the key reasons we are interested in releasing the work for community collaboration, both in terms of quality and quantity.

4 Comparison with Existing Benchmarks

Given that we are proposing a new evaluation metric, it is sensible to compare how scores on the ParaLex dataset (using our suggested evaluation methodology) compare with existing metrics. In this comparison, we deal with English only, as it is the language with the highest number of established evaluation benchmarks. Although the purpose of this paper is not to compare learning algorithms, we take sets of semantic word vectors trained in different ways and inspect how the benchmarks correlate with each other from a high-level point of view: the Word2Vec Skip-Gram GoogleNews embeddings [21], the Senna vectors [8], the fastText English vectors [5] and a set of the GloVe Wikipedia vectors [22].

Table 1. English language evaluation benchmarks calculated on a range of notable word vectors.

Model	MEN	MTurk	SimLex	WS353	Google	MSR	BLESS	Battig	Avg.	ParaLex
SkipGram	0.74	0.67	0.44	0.70	0.40	0.71	0.80	0.41	0.63	0.64
GloVe	0.74	0.63	0.37	0.52	0.72	0.62	0.82	0.41	0.60	0.92
Senna	0.79	0.70	0.45	0.64	0.61	0.82	0.84	0.47	0.66	0.81
fastText	0.57	0.58	0.27	0.43	0.11	0.15	0.56	0.39	0.44	0.6

Table 1 summarizes the results of the various word vectors over the various benchmarks. The table is adapted and extended from Jastrzebski et al. (2017) using the Python package available at `https://github.com/kudkudak/word-embeddings-benchmarks/`.

As we noted in section 1, current conclusions about correlation of different metrics with downstream tasks are chaotic. If anything, the results in Table 1 underline the lack of concrete conclusions possible. Scores on the ParaLex correlate with few of the other metrics and do not correlate with average performance on all other metrics. What is more, few of the established metrics correlate with each other.

5 Example ParaLex Results

One of the principal aims of our evaluation schema is its applicability on a multilingual scale. To this end, the results presented in section 4 are tangential to the principal aim of this project. Table 2 presents calculated scores of three established word embedding packages in all languages currently available for each: fastText [5], Polyglot [1] and ConceptNet Numberbatch[2] [26].

6 Conclusions and Future Development

This paper introduced ParaLex, a multilingual dataset for evaluating (distributional) semantic models. To the best of our knowledge, this is the only resource of its kind for evaluating semantic models across such a large number of languages. We release our dataset on GitHub and hope others will be able to make use of it. We also invite community contribution, both in terms of adding new, relevant clusters and ensuring that existing clusters are of good quality and consistency.

Acknowledgments Jussi Karlgren's work was partially supported by a VINNMER Marie Curie grant from VINNOVA, the Swedish Governmental Agency for Innovation Systems.

References

1. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed word representations for multilingual nlp. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. pp. 183–192. Association for Computational Linguistics, Sofia, Bulgaria (August 2013)
2. Avraham, O., Goldberg, Y.: Improving reliability of word similarity evaluation by redesigning annotation task and performance measure. In: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP. pp. 106–110. Association for Computational Linguistics, Berlin, Germany (2016)

[2] N.B: The Numberbatch project only provides a single set of vectors for Malay and Indonesian under the MS code. In the same package, Tagalog is indexed under FIL (Filipino) and for Croatian, the only vectors available are trained on data mixed with Bosnian and Serbian, indexed under SH.

3. Baroni, M., Lenci, A.: How we BLESSed distributional semantic evaluation. In: Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics. pp. 1–10. Association for Computational Linguistics, Edinburgh, Scotland (2011)
4. Batchkarov, M., Kober, T., Reffin, J., Weeds, J., Weir, D.: A critique of word similarity as a method for evaluating distributional semantic models. In: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP. pp. 7–12. Association for Computational Linguistics, Berlin, Germany (2016)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), https://transacl.org/ojs/index.php/tacl/article/view/999
6. Che, X., Ring, N., Raschkowski, W., Yang, H., Meinel, C.: Traversal-free word vector evaluation in analogy space. In: Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP. pp. 11–15. Association for Computational Linguistics, Copenhagen, Denmark (September 2017), http://www.aclweb.org/anthology/W17-5302
7. Chiu, B., Korhonen, A., Pyysalo, S.: Intrinsic evaluation of word vectors fails to predict extrinsic performance. In: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP. pp. 1–6. Association for Computational Linguistics, Berlin, Germany (2016)
8. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**, 2493–2537 (2011)
9. Faruqui, M., Dyer, C.: Community evaluation and exchange of word vectors at wordvectors.org. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Baltimore, Maryland, USA (June 2014)
10. Faruqui, M., Tsvetkov, Y., Rastogi, P., Dyer, C.: Problems with evaluation of word embeddings using word similarity tasks. In: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP. pp. 30–35. Association for Computational Linguistics, Berlin, Germany (2016)
11. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proceedings of the 10th international conference on World Wide Web. pp. 406–414. Association for Computing Machinery, Hong Kong, Hong Kong (2001)
12. Gerz, D., Vulić, I., Hill, F., Reichart, R., Korhonen, A.: SimVerb-3500: A large-scale evaluation set of verb similarity. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2173–2182. Association for Computational Linguistics, Berlin, Germany (2016)
13. Ghannay, S., Favre, B., Esteve, Y., Camelin, N.: Word embedding evaluation and combination. In: Proceedings of the 10th Edition of the Language Resources and Evaluation Conference. pp. 23–28. ELRA, Portoroz, Slovenia (2016)
14. Gladkova, A., Drozd, A.: Intrinsic evaluations of word embeddings: What can we do better? In: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP. pp. 36–42. Association for Computational Linguistics, Berlin, Germany (2016)
15. Gladkova, A., Drozd, A., Matsuoka, S.: Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In: Proceedings of NAACL-HLT. pp. 8–15. Association for Computational Linguistics, San Diego, California, USA (2016)

16. Hill, F., Reichart, R., Korhonen, A.: Simlex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics (2016)
17. Jastrzebski, S., Lesniak, D., Czarnecki, W.M.: How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks. CoRR **abs/1702.02170** (2017)
18. Karlgren, J., Callin, J., Collins-Thompson, K., Gyllensten, A.C., Ekgren, A., Jurgens, D., Korhonen, A., Olsson, F., Sahlgren, M., Schütze, H.: Evaluating learning language representations. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 254–260. Springer International Publishing (2015)
19. Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review **104**(2), 211 (1997)
20. Linzen, T.: Issues in evaluating semantic spaces using word analogies. In: Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP. pp. 13–18. Association for Computational Linguistics, Berlin, Germany (2016)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR (2013)
22. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014)
23. Rogers, A., Drozd, A., Li, B.: The (too many) problems of analogical reasoning with word vectors. In: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017). pp. 135–148. Association for Computational Linguistics, Vancouver, Canada (August 2017), http://www.aclweb.org/anthology/S17-1017
24. Santus, E., Yung, F., Lenci, A., Huang, C.R.: EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In: Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015). pp. 64–69. Association for Computational Linguistics and Asian Federation of Natural Language Processing, Beijing, China (2015)
25. Schnabel, T., Labutov, I., Mimno, D.M., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: EMNLP. pp. 298–307. Association for Computational Linguistics, Lisbon, Portugal (2015)
26. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge (2017), http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972
27. Turney, P.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. Machine Learning: ECML 2001 pp. 491–502 (2001)

Table 2. Multilingual scores on the ParaLex dataset for fastText, Polyglot and ConceptNet Numberbatch packages.

Code	Language	fastText	Polyglot	ConceptNet
AR	Arabic	41%	53%	54%
AZ	Azerbaijani	39%	30%	38%
BG	Bulgarian	51%	52%	74%
BN	Bengali	41%	28%	—
CA	Catalan	54%	43%	86%
CS	Czech	44%	47%	79%
DA	Danish	46%	52%	89%
DE	German	47%	71%	51%
EL	Greek	52%	41%	56%
EN	English	85%	80%	51%
ES	Spanish	53%	64%	63%
ET	Estonian	25%	37%	80%
FA	Persian	46%	42%	64%
FI	Finnish	42%	43%	63%
FR	French	52%	47%	44%
HE	Hebrew	39%	52%	78%
HI	Hindi	55%	42%	59%
HR	Croatian	44%	40%	59%
HU	Hungarian	49%	64%	90%
ID	Indonesian	62%	63%	—
IS	Icelandic	32%	21%	84%
IT	Italian	67%	57%	69%
JA	Japanese	44%	24%	66%
JV	Javanese	40%	30%	—
KO	Korean	27%	36%	86%
LT	Lithuanian	45%	46%	69%
LV	Latvian	41%	20%	78%
MS	Malay	57%	47%	90%
NL	Dutch	49%	55%	77%
NO	Norwegian	50%	66%	80%
PL	Polish	56%	46%	83%
PT	Portuguese	60%	50%	61%
RO	Romanian	45%	47%	55%
RU	Russian	59%	62%	73%
SK	Slovak	42%	41%	94%
SL	Slovenian	39%	39%	81%
SQ	Albanian	28%	18%	38%
SV	Swedish	34%	53%	90%
SW	Swahili	34%	23%	62%
TH	Thai	30%	29%	69%
TL	Tagalog	36%	21%	62%
TR	Turkish	48%	62%	79%
UK	Ukrainian	37%	31%	62%
UR	Urdu	34%	20%	53%
VI	Vietnamese	12%	12%	30%
ZH	Chinese	48%	54%	61%