Finnish resources for evaluating language model semantics

Viljami Venekoski
National Defence University
Helsinki, Finland
venekoski@gmail.com

Jouko Vankka
National Defence University
Helsinki, Finland
jouko.vankka@mil.fi

Abstract

Distributional language models have consistently been demonstrated to capture semantic properties of words. However, research into the methods for evaluating the accuracy of the modeled semantics has been limited, particularly for less-resourced languages. This research presents three resources for evaluating the semantic quality of Finnish language distributional models: (1) semantic similarity judgment resource, as well as (2) a word analogy and (3) a word intrusion test set. The use of evaluation resources is demonstrated in practice by presenting them with different language models built from varied corpora.

1 Introduction

In the spirit of the distributional hypothesis stating that semantically similar words appear in similar contexts (Harris, 1954), distributional language models of recent years have successfully been able to capture semantics properties of words given large corpora (e.g., Mikolov et al., 2013a; Pennington et al., 2014). However, there are only few resources for evaluating the accuracy or validity of language models, particularly for less-spoken languages, due to language dependence of such resources (Leviant and Reichart, 2015). Further, a single good resource may not be sufficient due to the complexity of semantics; performance in an intrinsic evaluation task may not predict performance in extrinsic downstream language technology applications (Chiu et al., 2016). Therefore, the evaluation of semantics should be based on a variety of tasks, estimating different semantic phenomena (Baroni et al., 2014).

With respect to language models, two distinct measures of semantic quality can be identified: validity and completeness (Lindland et al., 1994). The latter is dependent on the underlying corpus because a language model can only represent linguistic units which have been present in its training data. While it is possible for some models to infer the meaning of novel input, the inference can be considered an additional training step of the model and thus an extension of the training corpus. Completeness is also likely to affect the validity of a model; given the distributional hypothesis (Harris, 1954), more encompassing knowledge about the possible contexts of words results in more accurate knowledge of their semantics. In this study, the lack of completeness is estimated only by presenting a rate of out-of-vocabulary (OOV) words for each separate evaluation resource, but not investigated further.

The aim of this research is to present scientist and practitioners working with Finnish tools to evaluate their language models with respect to semantics.[1] While most research compares the performance of models to that of humans, we also present effortlessly extensible tools requiring no human annotation. Finally, baseline results for the evaluation methods are reported, utilizing varied corpora and language model architectures.

2 Materials and Methods

2.1 Language models

The distributional language models used in this research are constructed using word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2016) software. These model architectures have been used to produce vector representations of words, known as word embeddings, efficiently from large corpora, with the vectors yielding intuitively semantic properties (Mikolov et al., 2013b; Baroni et al., 2014). The word embeddings have been used in a vari-

[1] The evaluation resources are available online at github.com/venekoski/FinSemEvl.

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 231–236,
Gothenburg, Sweden, 23-24 May 2017.

ety of semantics-incorporating downstream applications such as sentiment analysis and text generation (see e.g., Brigadir et al., 2015; Karpathy and Fei-Fei, 2015; Bansal et al., 2014).

The models are constructed utilizing the default hyperparameters as set out by the authors of each model, except for the minimum word frequency which is set to 5 for each model. Both Continuous Bag of Words (CBOW) and Skip-gram (SG) architectures of word2vec and fastText are used. It should be noted that these parameters may not be optimal, particularly for Finnish (Venekoski et al., 2016), and tuning of the parameters would likely lead to better results.

2.2 Corpora and pre-processing

The language models are created based on four different publicly available Finnish corpora. These include the Suomi24 (Aller Media Oy, 2014) and Ylilauta (Ylilauta, 2011) corpora of social media discussions, as well as corpora derived from a Wikipedia dump (Wikimedia Foundation, nd) and all Finnish language Project Gutenberg texts (Project Gutenberg, nd). No pre-processing other than lowercasing of characters was conducted on the data. The descriptives of the corpora are reported in Table 1.

Corpus	Tokens	Unique tokens
Suomi24	2834M	31508K
Ylilauta	30M	524K
Wikipedia	79M	2747K
Gutenberg	72M	2034K

Table 1: Corpora and their sizes after tokenization.

2.3 Word similarity resource

Arguably, the standard for evaluating semantic accuracy of language models is using word similarity resources. These typically comprise of a set of word pairs, each having a human-rated similarity score. The human ratings are then correlated with similarity scores produced by a computational language model. Among the most utilized resources are WordSim-353 (Finkelstein et al., 2001), MEN (Bruni et al., 2012), and SimLex-999 (Hill et al., 2015). However, the resources differ in what they quantify; the instructions of WordSim-353 lead its respondents to rate association between words (Agirre et al., 2009), whereas the instructions of SimLex-999 guided the subjects to evaluate similarity between words specifically (Hill et al., 2015). Notably, performance in SimLex-999 predicts the performance in downstream applications, unlike most intrinsic evaluation benchmarks (Chiu et al., 2016).

Similarity judgment resources cannot be used cross-lingually by translating a resource to another language and using the scores of the original resource to evaluate cross-language models (Leviant and Reichart, 2015; but see Agirre et al., 2016). Thus, in order to evaluate Finnish language models, a new similarity resource based on SimLex-999 (henceforth SL999) was constructed. Following the same instructions as in SL999, an online survey was conducted in which respondents were asked to rate the similarity of pairs of two words on a scale of 0 to 10, where 0 meant no similarity between the words while 10 meant that the words were synonymous. The survey consisted of 300 word pairs from SL999 which were translated to Finnish. The chosen words each had a single unambiguous sense in both Finnish and English, hence excluding homographic words. This was to ensure that the Finnish participants would rate words denoting senses most similar to their English counterparts, allowing cross-lingual comparisons. The translations were agreed upon by two fluent bilingual researchers. The inflectional form of the Finnish words was singular nominative for nouns and adjectives and first infinitive for verbs. Finally, the set was randomly reduced to 300 pairs to reduce survey fatigue of the respondents. The presentation order of word pairs was randomized for each respondent. The survey was conducted online and the respondents recruited through social media. Only native Finnish speakers were instructed to fill out the survey. To filter out outliers, the exclusion criterion of SL999 was followed: the respondents whose answers' average Spearman correlation with all other respondents' answers deviated from the mean of all such averages by more than one standard deviation were excluded. As a result, 4 out of 59 total respondents were excluded from the subsequent analyses. The resulting data set of similarity ratings for 300 Finnish word pairs as judged by 55 respondents will henceforth be called the FinnSim-300 (or FS300) data set.

To obtain a human performance benchmark, inter-annotator agreement was calculated as the average pairwise Spearman correlation between two human raters. The agreement was $\rho = .53$, and while lower than the agreement in SL999 (Hill

et al., 2015), $\rho = .67$, it can be considered sufficiently high as inter-annotator agreements can be relatively low in similar ambiguous tasks (Gamon, 2004; Strapparava and Mihalcea, 2007). The lower agreement in the Finnish resource and related greater variance in individual responses can be partially attributed to the fact that respondents in SL999 rated word pairs on a smaller scale of $1-7$ (which the researches extrapolated to a $0-10$ scale). The standard deviations of respondent similarity ratings for individual items in SL999 ranged from .34 to 2.18 (Hill et al., 2015), scoring notably higher compared to a range of .13 to 3.35 in the current Finnish survey.

More recently, it has been argued that the average correlation of one human rater with the average of all the other raters is a fairer measure for evaluating computational models performance compared to inter-rater agreement (Mrkšić et al., 2016). This score, gold standard agreement, was $\rho = .72$ in FS300, which is also more in line with the score in SL999, $\rho = .78$.[2] We consider this value as a point of comparison for language model evaluation. Should the correlation between the similarity scores of a language model and the similarity judgment resource exceed this number, the model can be considered to perform at a human level.

2.4 Analogies

Analogies have previously been used as a method for evaluating the semantic reliability of language models (see e.g., Bojanowski et al., 2016; Sun et al., 2016). Alongside word2vec model, its authors released an English language analogy test set, consisting of approximately 20 000 syntactic and semantic test units, each following the analogy A is to B what C is to D (Mikolov et al., 2013a). In the test task, a well-performing model is expected to estimate the correct word D given vectors of words A, B and C, by estimating the most similar word vector to that obtained from the linear operation $\mathbf{w}_B + \mathbf{w}_C - \mathbf{w}_A$. The test is correct if the most similar word is exactly that which has been determined by the test set. If there were no vector representation for one of the words in the analogy, the analogy determined incorrect. The overall percentage of correct analogies indicates the extent in which a model is able to capture known semantic relations.

[2]Reported by the authors at: http://www.cl.cam.ac.uk/%7Efh295/simlex.html.

The words the original English test set are not directly applicable to other language models, if translated, due to culture-specific terminology (e.g. US newspapers and sports teams). Thus, a small Finnish analogy test was created, consisting of 1037 analogies. The relation types in the analogy set were in part taken from the Google analogy set (capital-country, country-currency, female-male), but extended with other relation types as well (antonymic adjectives, orthogonal directions, hockey team-city, cardinal-ordinal number).

The semantics of models are highly reliant on the conceptual knowledge that is exhibited in the data. The human authors of the underlying corpus may have distorted conceptual knowledge compared to the curated analogy test sets. For instance, an individual may think that the capital of Australia is Sydney and consequently produce utterances corresponding to this proposition. Thus, even if a model would be able to perfectly capture the conceptual information of said individual, the model would fail at an analogy task utilizing capital-country relations. Therefore, the analogy task is not only an evaluatory tool for the semantic validity of a language model but also for the conceptual validity of the corpus. If only the former evaluation is desired, effort should be put onto making the analogy test sets such that they contain unambiguous, uncontroversial, common knowledge factual relationships, where the to-be-guessed word (notated word D above) is the only correct alternative in the vocabulary.

2.5 Word intrusion

Word intrusion task (also known as *odd-one-out* or *oddity task*) is a traditional experimental paradigm in psycholinguistic research where the subject is instructed to choose a word which is semantically incompatible with rest of the words in a list (see e.g., Albert et al., 1975; Campbell and Sais, 1995). More recently, the paradigm has been used in machine learning literature to evaluate the semantic coherence of topic models (Chang et al., 2009). In this setting, a list of n words (we call this an intrusion set) is created, out of which $n-1$ words are taken from one topic, constructed by the topic model, and one outlier word is taken from another topic. Human subjects are asked to point out the outlier word, and if they agree with the topic model partition, the model is considered coherent.

The aim here is to utilize the intrusion task but

Task	Corpus	OOV	*GloVe*	*word2vec*		*fastText*	
				CBOW	SG	CBOW	SG
Similarity judgments	Suomi24	0.00%	.2381	.2431	.3070	.3724	**.3788**
	Ylilauta	1.00%	.0823	.1689	.1876	**.2605**	.2379
	Wikipedia	0.67%	.1385	.1460	.1855	**.2780**	.2121
	Gutenberg	10.67%	.2312	.2583	.2953	**.3430**	.3323
Analogies	Suomi24	0.00%	.1861	.1302	.1948	.0366	**.1986**
	Ylilauta	23.14%	.0897	.0984	**.1485**	.0492	.0916
	Wikipedia	0.00%	**.4330**	.2507	.3655	.1543	.4098
	Gutenberg	40.12%	.0540	.1138	**.1340**	.0569	.0887
Word intrusion	Suomi24	52.09%	.3983	.7227	**.8297**	.6081	.8288
	Ylilauta	81.89%	.3449	.5846	**.7016**	.4805	.6944
	Wikipedia	51.17%	.5484	.8116	**.9207**	.6805	.8825
	Gutenberg	85.75%	.2938	.4272	.5329	.4620	**.5901**

Table 2: Performance of different language models in the three presented evaluation tasks. The scores of the best performing models for each corpus in each evaluation task are marked in bold.

turn attention away from evaluating coherence of topics and towards evaluating the language model itself. We conduct the same task but manually construct the intrusion sets from words which are known, a priori, to belong to specified topics. The topics used in this research comprised of lists of articles from Finnish language Wikipedia. In total, 16 lists were extracted, including lists of sports, illnesses, minerals, and professions, among others. The lists contained 4127 unique items in total. In order to evaluate the language model and not the underlying corpus, only words which had a vector representation in the language model under evaluation were included in the intrusion sets.

Following (Chang et al., 2009), the size of the intrusion set was set to 6 words, where 5 words are randomly sampled from word list A and one outlier word from list B. The intrusion task is conducted 10 000 times with each ordered pair of the given lists. This was done to increase the task's reliability by reducing effects arising from random sampling of words from variable-length lists. A models score is the overall percentage of correctly determined intruder words.

3 Results

To demonstrate the evaluation methods in effect, results on multiple distributional language models are presented in Table 2. Out of vocabulary rates are reported for each task given a corpus.[3] The Gutenberg corpus has the highest OOV rate in all tasks, suggesting that these evaluation sets function best with models built from contemporary corpora.

[3]OOV words were excluded from similarity and intrusion tasks but included and counted as errors in the analogy task.

While the performance of different models is varied between different tasks and corpora, some trends can be observed. The word2vec Skip-gram and fastText models appear to produce better results compared to GloVe and word2vec-CBOW models. Conceptual relations as measured by the analogy task are best captured from Wikipedia corpus, while the large social media corpus of Suomi24 achieves the best correspondence with human similarity judgments.

4 Conclusions

In this study, Finnish language resources for evaluating semantic accuracy of language models were presented. Such resources are necessary for optimizing language model construction and they give researchers quantifiable estimates for the extent in which models are able to capture meaning from data. The resources constructed include a similarity judgment resource FinnSim-300, an analogy test set, and a word intrusion test set. Future research is encouraged to expand and adapt the resources because corpus and domain-specific test sets are likely to be more appropriate for most evaluations. While the presented methods serve as a starting point for evaluating semantic accuracy, a thorough discussion on what aspects of semantics can be reliably quantified is needed. Good performance in evaluation tasks provides basis for the claim that computational models can indeed be valid and reliable models of semantics.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Eneko Agirrea, Carmen Baneab, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Rada Mihalceab, German Rigaua, Janyce Wiebef, and Basque Country Donostia. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. *Proceedings of SemEval*, pages 497–511.

Martin L Albert, Avinoam Reches, and Ruth Silverberg. 1975. Associative visual agnosia without alexia. *Neurology*, 25(4):322–326.

Aller Media Oy. 2014. The Suomi 24 Corpus. May 14th 2015 version, retrieved October 27, 2016 from http://urn.fi/urn:nbn:fi:lb-201412171.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *ACL (2)*, pages 809–815.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Igor Brigadir, Derek Greene, and Pádraig Cunningham. 2015. Analyzing discourse communities with distributional semantic models. In *Proceedings of the ACM Web Science Conference*, page 27. ACM.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Ruth Campbell and Efisia Sais. 1995. Accelerated metalinguistic (phonological) awareness in bilingual children. *British Journal of Developmental Psychology*, 13(1):61–68.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 288–296.

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 1–6. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM.

Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR*, abs/1508.00106.

Odd Ivar Lindland, Guttorm Sindre, and Arne Solvberg. 1994. Understanding quality in conceptual modeling. *IEEE software*, 11(2):42–49.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Project Gutenberg. n.d. Retrieved October 27, 2016 from www.gutenberg.org.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Semantic regularities in document representations. *arXiv preprint arXiv:1603.07603*.

Viljami Venekoski, Samir Puuska, and Jouko Vankka. 2016. Vector space representations of documents in classifying finnish social media texts. In *Proceesings of the 22nd International Conference on Information and Software Technologies, ICIST 2016*, pages 525–535. Springer.

Wikimedia Foundation. n.d. Wikipedia – the free encyclopedia. 2016-10-20 dump, retrieved October 27, 2016 from https://dumps.wikimedia.org/fiwiki/20161020/.

Ylilauta. 2011. Ylilauta Corpus. March 4 2015 version, retrieved October 27, 2016 from http://urn.fi/urn:nbn:fi:lb-2015031802.