Is Similarity Visually Grounded? Computational Model of Similarity for the Estonian language

Claudia Kittask
Institute of Computer Science
University of Tartu
Tartu, Estonia
claudiakittask@gmail.com

Eduard Barbu
Institute of Computer Science
University of Tartu
Tartu, Estonia
eduard.barbu@ut.ee

Abstract

Researchers in Computational Linguistics build models of similarity and test them against human judgments. Although there are many empirical studies of the computational models of similarity for the English language, the similarity for other languages is less explored. In this study we are chiefly interested in two aspects. In the first place we want to know how much of the human similarity is grounded in the visual perception. To answer this question two neural computer vision models are used and their correlation with the human derived similarity scores is computed. In the second place we investigate if language influences the similarity computation. To this purpose diverse computational models trained on Estonian resources are evaluated against human judgments.

1 Introduction

Various disciplines and research communities are interested in the study of similarity: Philosophy, Psychology, Computational Linguistics, Semantic Web and the Linked Data communities, to name a few. For example, to integrate heterogeneous semantic resources the Linked Data and Semantic Web communities estimate the degree of similarities between the concepts in these resources (Euzenat and Shvaiko, 2013; Harispe et al., 2015).

In Computational Linguistics researchers build computational models of similarity. To test them, the correlation between the human similarity scores and the scores assigned by the computational models is calculated. It is assumed that the best computational models predict better the human judge scores.

However, earlier studies of similarity suffered from a drawback: they do not distinguish between the relations of similarity and association. In psychology, for example, the distinction between these two notions is well understood. The association between two concepts is defined as the propensity of a subject to activate a representation of the second concept when the first concept is presented. In contrast, the similarity is defined as the proximity of two mental representations. In the Gestalt psychology for example (Wertheimer, 1938) the similarity is seen as the principle of organization of objects in perceptual groups. The concepts **cup** and **tea** are associated but not similar: there is no perceptual principle to group together an object like a cup and a liquid like tea. However, the objects denoted by the concepts **apple** and **pear** are perceptually similar. To remedy this problem SimLex-999 (Hill et al., 2015), a genuine data similarity set containing human judge similarity and concreteness scores for 999 English word pairs, has been built.

An interesting distinction is that between the surface similarity and deep similarity (Vosniadou and Ortony, 1989). The surface similarity is perceptually grounded and it is used in categorization. In contrast, the deep similarity is related to deeper properties not readily accessible to perception. A question we study is: How much of the similarity is grounded in the perceptual properties? In this research the degree of similarity grounded in visual properties is estimated by computer vision models.

If for the English language computational models of similarity have been implemented and evaluated, this is not the case for other languages. In particular, for the Estonian language there is no human annotated set that reflects the true similarity. We translate the SimLex-999 into Estonian and evaluate computational models of similarity based on Estonian language resources: corpora, taxonomies and lexical ontologies.

The rest of the paper is organized as follows. The next section puts our research in context and then the EstSimLex-999 set, the SimLex-999 set translated into Estonian, is presented 3. Section 4 discusses the families of models for computing the similarity between the word pairs in EstSimLex-999. Section 5 presents and discusses the results. In particular, we answer how similarity is influenced by language and quantify the power of the computer vision models to capture the similarity for concrete concepts. The paper ends with the conclusions.

2 Related Work

Felix Hill and coauthors (2015) undertook an extensive discussion of the concept of similarity in Computational Linguistics, introduced the genuine similarity set SimLex-999 and computed the correlations between the similarity measures of corpus based computational models and the human judge scores. Closely following this line of research Ira Leviant and Roi Reichart (2015) translated SimLex-999 in Italian, German and Russian and collected similarity scores from native speakers. They compute correlations between the human judgments and Vector Space Models (VSM) in a multilingual setting.

In the subsequent research the authors improve the similarity computational models and boost the correlation coefficient with the human judgments. For example, Schwartz et al. (2015) learn a word level representation based on symmetric patterns that achieves a Spearman correlation of 0.517 with SimLex-999. An interesting work belongs to Faruqui and Dyer (2015), who used non-distributional word representations derived form Princeton WordNet, FrameNet and Penn Treebank to reach a Spearman correlation of 0.58 with SimLex-999. Hybrid models (Recski et al., 2016),combining features from lexical ontologies and word embeddings, seem to be even better (Spearman Correlation 0.76).

In this work we were not interested in obtaining the best correlation between the computational models and the human judgments. That will be the topic of a future work. Instead, we were concerned with three problems. First, we are interested in how much of similarity is grounded in the visual features, that is, how much of the similarity is surface similarity. By evaluating the similarity using computational vision models we contribute to a better understanding of the notion of similarity itself. Second, we ask how the traditional models derived from Estonian corpora and lexical ontologies correlate with the judgments of native Estonian speakers. In this way we extend the similarity study to other language, a less explored one, yet an interesting one. Third, we study if our computational models trained on Estonian data predict better the EstSimLex-999 scores or the SimLex-999 scores. More precisely we want to know if the language influences the similarity judgments.

3 EstSimLex-999

To translate SimLex-999 the Google Translation API and a bilingual English-Estonian dictionary containing 87665 entries have been used, obtaining rough Estonian equivalents. A native Estonian speaker has chosen the correct translations. If an English word in a similarity pair is ambiguous, the sense that makes the pair more similar is preferred. Finally, after correction and the discussion with an Estonian linguist we have produced the similarity set referred from now on as EstSimLex-999. When translating, we have been careful to preserve the part of speech of the English concepts. This makes the comparison between the computational models of similarity for English and Estonian easier. Nevertheless, due to cultural and linguistic differences some English similarity pairs were hard to translate. For example, the English pair (taxi, cab) was translated as (taksi, takso) even if the second term of the Estonian pair is not widely used. Another example is the pair (supper, dinner). The Estonian culinary tradition does not distinguish between the two concepts, therefore we have translated the pair with the synonymous words (õhtusöök, õhtueine). Please, notice, that for many non-British native English speakers the words *supper* and *dinner* are also synonymous. Some translations would have been more accurate using multiwords, but we abide by the original requirement that the similarity pairs should contain single words only. Overall, we have produce an accurate translation of the English original SimLex-999 set preserving the distribution of the part of speeches and satisfying the demand that the word pairs should not contain multiple words.

Four native Estonian speakers have rated the degree of similarity between each of the 999 pairs.

The rating instructions are the same as in the original study (Hill et al., 2015). These instructions do not attempt to define what similarity is, but rather clarify the concept contrasting it with association, and comparing it with synonymy. The inter-annotator agreement was computed as the average of pairwise Spearman correlations between the scores of all raters. The overall agreement is 0.766. A direct comparison with the correlation coefficient computed in the English study (0.67) is not possible because the number of annotators is different. At this stage we were not interested in recruiting many annotators through platforms like Mechanical Turk, but rather in gaining insights into human similarity judgments by direct discussion with the annotators. In any case, recruiting a comparable number of Estonian speakers is unlikely, as this language is natively spoken by less than 1 million people. The main thing noticed is that there are few pairs of adjectives and verbs highly rated in English but with a low score in Estonian. For example, the English verb pair (appear, attend) has a score of 6.28 in SimLex-999 and its Estonian translation (ilmuma, osalema) has a score equal to 0.5.

4 Models for Similarity Computations

Three families of similarity models are evaluated : distributional models, semantic network models and computer vision models.

The distributional models are an implementation of John Rupert Firth's hypothesis "You shall know a word by the company it keeps" (Firth, 1961) which basically states that words that have similar meanings appear in comparable syntagmatic contexts. Nowadays, the most advanced distributional models are the neural word embeddings.

The second family of models derive the semantic similarity from the taxonomic structure of semantic networks. The IS-A relation induces the inheritance of the properties.The above mentioned concepts, *apple* and *pear*, are similar because they inherit all the properties from their superordinate concept (fruit). Unlike the distributional models, the semantic networks tells us also why the concepts are similar.

The third family of models are the computer vision models. The similarity between two concepts is the distance between their image representations. Because of the visual nature of this similarity the computer vision models work best for concepts representing concrete objects.

In what follows we will briefly describe the models tested.

1. **Word2Vec**. Word2Vec is a distributional model (Mikolov et al., 2013) implemented as a two layer neural network. If two words appear in similar contexts in a corpus the network will output embedding vectors, known as neural vector embeddings, which are close in the embedding space. Word2Vec computes the neural vector embeddings either predicting the target word from the context (this method is known as continuous bag of word (CBOW)) or as the target context from the word (this method is know as Skip Gram).

2. **SenseGram**. SenseGram (Pelevina et al., 2016) is not a distributional model per se, but a method to obtain word senses from word embeddings. This word discrimination method takes as input word embeddings (like those generated by Word2Vec or any other distributional model) and clusters them. The induced word senses correspond to the clusters of word embeddings.

3. **Path Similarity Measures**. The path similarity measures exploit the graph structure of semantic networks to find similarities between concept pairs. We have explored various similarity measures like Leacock-Chodorow similarity (Leacock and Chodorow, 1998).

4. **Autoencoders**. Autoencoders are deep neural networks which learn to reconstruct the input. In the reconstruction process one of the autoencoder layers contains less nodes than the input layer, thus forcing the network to learn a lower level representation of the input. The idea behind using the autoencoders is that the sparse representations learned when encoding similar concepts will be close in the embedding space.

5. **Pretrained Convolutional Neural Networks**. Convolutional Neural Networks (CNN) are deep neural networks architectures suitable for extracting patterns from images. Inspired by experiments in neuroscience (Hubel and Wiesel, 1959), CNN's

first layers train convolution filters to detect low-level features of an image like lines and corners. Higher network levels combine the low level-features to find high-level image features roughly corresponding to the human language semantic descriptions of the objects. For example, they might detect parts, like the wheels or the hood of a car. When CNN are trained on big databases of classified images the semantic representations of the concrete concepts can be "read" from the deeper network levels. These representations are then used to compute concept similarity.

5 Results

First the results for the distributional models are shown, then the results for the semantic network models will be presented. Finally, the results for the neural computer vision models will be shown. The correlation coefficients between the scores assigned by the computational similarity models for interesting subsets (e.g. abstract and concrete concepts) and the two similarity sets are also computed. In the tables in this section the SimLex-999 is abbreviated as SL-999 and the EstSimLex-999 as ESL-999.

5.1 Distributional Models

When evaluating the **Word2Vec** and **SenseGram** models, if the embedding vectors corresponding to the words in the SimLex-999 or EstSimlex-999 are missing, the word-pair is eliminated. The model word similarity score is computed as cosine similarity between the vector embeddings corresponding to the words in the each word pair. Pearson (r), Spearman (ρ), and Kendall (τ) correlations are calculated between EstSimLex-999 and Simlex-999 human judge scores and the model word similarity scores. The word embeddings were trained on Estonian monolingual corpora and the Estonian Wikipedia. The following word embeddings have been used:

- **EA word embeddings**. 9 Skip-Gram and 20 CBOW models, with different parameter settings, were trained on the lemmatized version of etTenTen corpus of Estonian Web [1] by Eleri Aedma. Word senses were induced from the traditional word embeddings using SenseGram. SenseGram finds 1.6 senses/concept, with about 300 word pairs having more than one sense. For the ambiguous word pairs (where at least one of the words in the pair has more than one sense) the word sense that maximizes the cosine similarity score of a word pair is evaluated.

	SL-999			**ESL-999**		
Model	r	ρ	τ	r	ρ	τ
cbow_1	.42	.42	.29	.46	.47	.33
sg_2	.37	.36	.24	.41	.42	.3
cbow_3	.33	.33	.23	.33	.34	.24

Table 1: The results for the best three distributional models

- **Estnltk pretrained word embeddings**. Estnltk (Orasmaa et al., 2016) contains 8 word pretrained embeddings. 4 of them are trained with the CBOW method and the other 4 were trained with the Skip-Gram method, on the raw and lemmatized versions of the Estonian Reference Corpus (Kaalep et al., 2010). The Estonian Reference Corpus is a 1.3 billion word corpus, crawled from the web, containing mainly newspaper text.

- **Facebook pretrained word embeddings**. The Facebook word embeddings (Bojanowski et al., 2017) have been trained with CBOW method on 294 language versions of Wikipedia.

The distributional models evaluate on average 985 word pairs. Each CBOW and Skip Gram model has 4 meta-parameters : the number of dimensions, the window size, the minimum count threshold and the number of iterations. Due to consideration related to space we only present the best three results in the table 5.1. The whole set of results for the 67 distributional models trained and all the figures and the tables in this paper are available online linked from our github repository. [2]. The best model on the first row in the table 5.1, for example, has been trained with the 300 dimensions, a window size equal to 1, the minimum count threshold being 10, and 20 iterations.

In the first place one can notice that CBOW trained word embedding perform better than Skip-Gram trained word embeddings. Moreover, the correlation coefficients between EstSimLex-999

[1] DOI: 10.15155/1-00-0000-0000-0000-0012EL

[2] https://github.com/estsl/EstSimLex-999

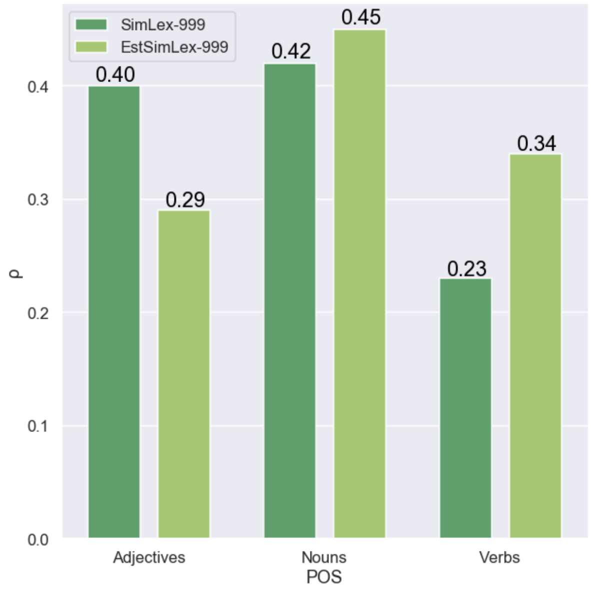

Figure 1: The average performance for POS-based subsets

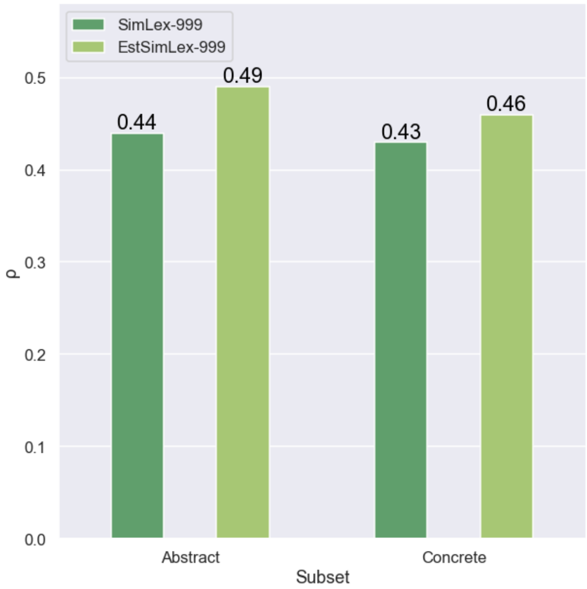

Figure 2: The average performance for the most concrete and abstract subsets

human scores and the model computed word similarity are higher than the correlation coefficients between SimLex-999 human scores and the model computed word similarity. The automatic sense discrimination had a negative influence on the results, as the SenseGram induced vector senses show a slight drop in performance over the traditional word-embeddings. The Estnltk trained word embeddings perform worse than EA word embeddings, but better than Estonian Facebook pretrained word embeddings.

Furthermore, we study if the part of speech category influences the strength of the correlation between human judgments and the distributional models. The model similarity scores between the words in the word pairs is the average similarity scores for all the distributional models.The correlation coefficients between the model scores for 666 noun pairs, 111 adjective pairs and 222 verb pairs and the human judgment scores is calculated.

As it can be seen in figure 1, the best (Spearman) correlation coefficients between the models and the similarity sets are obtained for the nouns. The (Spearman) correlation coefficient between the distributional models and the human judgments is higher for EstSimLex-999 set than for the original SimLex-999 set.

The correlation coefficients between the 250 most concrete word pairs and the 250 most abstract word pair and the distributional models have also been computed. The results presented in figure 5.1 show that, on average, the distributional models correlate better with the abstract human judgments scores and that the correlation coefficient is higher for the EstSimLex-999 set.

5.2 Semantic Network Models

The similarity between the concepts corresponding to the words in EstSimLex-999 word is computed for two semantic networks: the Estonian Wordnet and a taxonomy derived from the Estonian Wikipedia.

As the Estonian Wordnet lists multiple senses for words, a disambiguation procedure to select the most likely sense is implemented. The Cartesian product between the word senses in the semantic network corresponding to the words in the EstSimLex-999 is generated. Thus we obtain a set of word-sense pairs. Subsequently, as explained below, a similarity scores is assigned to each word-sense pair in this set. The word sense pair that maximizes the similarity score is chosen. This procedure of mapping the words onto a semantic network is very effective, obtaining over 90 percent precision for the Estonian Wordnet (Barbu et al., 2018) .

Three similarity measures between the semantic network concepts have been computed: path similarity (PS), Leacock & Chodorow similarity (LC) (Leacock and Chodorow, 1998) and Wu &

	SL-999			ESL-999		
	r	ρ	τ	r	ρ	τ
PS	.47	.47	.35	.54	.52	.39
LC	.36	.36	.26	.41	.43	.31
WuP	.41	.45	.32	.49	.53	.39

Table 2: The results for the Estonian Wordnet

	SL-999			ESL-999		
	r	ρ	τ	r	ρ	τ
PS	.32	.31	.22	.37	.34	.24
LC	.31	.3	.21	.35	.34	.28
WuP	**.39**	**.37**	.28	.4	**.37**	.27

Table 3: The results for the Wikipedia Page Taxonomy

Palmer similarity (WuP) (Wu and Palmer, 1994). Pearson (r), Spearman (ρ), and Kendall (τ) correlations were calculated between EstSimLex-999 and SimLex-999 human judge scores and the network similarity scores for the disambiguated word pairs.

The Estonian Wordnet is an ongoing effort, it pursues roughly the same organization principles as Princeton WordNet (Miller et al., 1990), and it is manually built by a group of linguists. The version used in this study contains approximately 85.000 synsets. The above described disambiguation procedure maps approximately 770 word pairs onto the Estonian Wordnet. The results corresponding to the Estonian Wordnet are in the table 2.

The best results are obtained for Wu & Palmer similarity measure and EstSimLex-999. This similarity measure considers the depth of the concepts in the semantic network hierarchy along with the depth of their Lowest Common Subsumer. Unlike the path similarity measure it favour the concepts that are deeper in the hierarchy. As in the case of the distributional models the EstSimLex-999 correlation scores are better than the SimLex-999 ones. A fact worth noticing is that the difference between the correlation scores for the two similarity sets is greater when we compute the similarity score based on the Estonian Wordnet structure instead of estimating the similarity using distributional models.

The taxonomy was extracted from the (Estonian) Wikipedia page text (Wikipedia Page Taxonomy) by the language technology research group at Università Roma Tre (Flati et al., 2016). The Wikipedia Page Taxonomy contains approximately 87000 concepts. We could map around 200 word pairs onto the Wikipedia Page Taxonomy. The results for this taxonomy are in Table 3.

The correlation coefficients between the similarity measures computed for the Wikipedia Page taxonomy and the human judgments scores are much lower than those computed before (with the Estonian Wordnet). Also, surprisingly and for the first time in this study, there is no statistically significant difference between the correlation coefficients computed with SimLex-999 human judgments scores and the EstSimLex-999 human judgment scores.

5.3 Computer Vision Models

Because the visual similarity is correlated with the level of concreteness of an object, the computer vision models are fed with word pairs where both words have a degree of concreteness higher than the a threshold equal to 4.8 . This criterion gives us 136 word pairs. We have downloaded, using Yandex image search engine 200 images for each word in the selected word pairs.

The first architecture trains a Convolutional Autoencoder (CAE) on the downloaded images. The encoder consists of 3 convolutional layers, each followed by a max-pooling layer. The decoder consists of 3 convolutional followed by upsampling layers. The similarity between two images is calculated as the cosine similarity between the corresponding encoder vectors. The similarity score for a word pair is the average score between all the images corresponding to the words in the pair.

The second architecture is the winner of the ImageNet 2015 competition. It is a CNN network architecture invented by Microsoft Research, called ResNet(He et al., 2016) (abbreviation for Residual Network). DNNs with many layers are difficult to train due to vanishing and exploding gradient problems. ResNet solves these problems with residual learning. The ResNet architecture comes in many variants, depending on the number of layers the network has. The widely employed ResNet-18 variant with 18 layers is used in this study. The network is pretrained on the ImageNet database (Deng et al., 2009) which contains over 1 million images classified under 1000 Princeton WordNet categories. Being trained on such a big

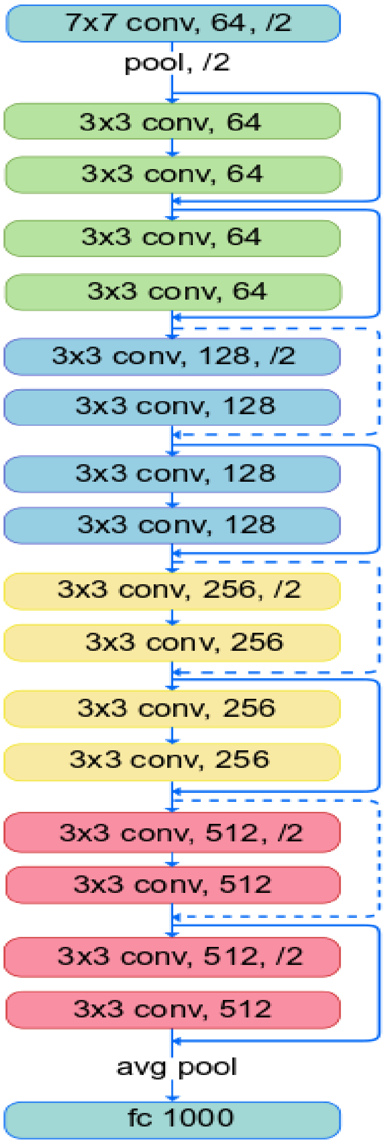

Figure 3: Architecture of ResNet-18

and diverse database of images, ResNet learns rich feature representations that help discriminate between images belonging to different categories.

The architecture of ResNet-18 network is presented in 5.3. As can be noticed the network repeats the same structure, and it ends with the average pool layer that feeds a fully connected layer. The probability that an image belongs to a category is computed by a softmax layer. The network is fed with the downloaded images corresponding to the words in the selected concrete set. The image representation learned by the network is read from the average pool layer. The score this visual model assigns to a word pair is the average cosine similarity score between the pair words image representations. As a way of example let's take the hypothetical word pair (cat, dog). The score assigned by the model to this pair is the average cosine similarity score between the ResNet-18 representations of the images corresponding to the words dog and cat.

Model	SL-999			ESL-999		
	r	ρ	τ	r	ρ	τ
CAE	.25	.28	.19	.17	.22	.15
RN18	.37	**.38**	.26	.34	**.38**	.27

Table 4: The results for the convolutional autoencoder (CAE) and ResNet-18 (RN18) architectures

The results for the two computer vision architecture are presented in Table 4. As expected, the rich visual features learned by the ResNet-18 architecture boost the results (the best results are bold marked in the table) for both similarity sets. The highest correlation coefficients are between the computer vision models and both similarity sets human judge scores.

For the three families of models three correlation coefficients have been computed. These coefficients do not induce a different order on the results, therefore the usage Spearman correlation coefficient in the previous studies is justified.

In general the correlation coefficients between all families of computational models and the human judge scores of EstSimLex-999 are better than the same correlation coefficients and the human judge scores of SimLex-999. This result shows that there is a slight language effect on the perception of similarity.

Regarding the magnitude of the Spearman correlation coefficient for the Estonian side of the equation, the distributional models show a moderate strength [3] of correlation with the human judge scores. This means that the distributional models do capture some of the notion of similarity between the words, but they also capture something else. The best Spearman correlation coefficient is obtained with the Estonian Wordnet (0.53), being better than the best correlation coefficient for distributional models (0.47), but still in the moderate range. It seems that the best predictor of human similarity is derived from a manually built resource containing clearly defined semantic relations.

Less than 40 percents of the similarity of the concrete concepts can be explained by the visual

[3] In the literature the strength is moderate if the coefficient is in the range 0.4-0.59.

semantic features when these features are computed from huge databases of images like ImageNet. Although the empirical evidence heavily depends on the quality of the ImageNet database this finding shows that other factors have significant weight in human similarity judgments. Maybe a case can be that these factors are the deep semantic features we have briefly mentioned in the introduction section. However, a definitive answer to what these factors are and how to account for them goes beyond this research.

6 Conclusions

In this study have addressed some aspects of the computational models of similarity applied to the Estonian language.

In the first place we have found that the neural visual models of similarity can explain a part of the similarity between the words representing concrete concepts. This invites the conclusion that deep similarity models might be involved in accounting for the unexplained part of similarity.

In the second place and differently from the finding in (Leviant and Reichart, 2015) an effect of the language on the similarity has been found. Unlike in that study the Estonian computational models of similarity better correlate with the word pair similarity score assigned by the Estonian subjects that with the scores assigned by English speaking subjects.

In the third place the best computational models are those derived from human built semantic networks. They are better than the neural distributional models but still they correlate moderately with the human judgments. This means that there is more to similarity than taxonomic similarity. On the other hand the Estonian Wordnet is still work in progress, therefore we cannot rule out that a more complete wordnet can boost the similarity scores. Unlike the original study (Hill et al., 2015) we have found that the word embedding computational models better correlate to the scores for nouns and not the adjectives. Intuitively, this result makes sense as nouns have richer mental representations than other morphological categories, therefore one expects that the similarity is better defined for nouns than for other parts of speech.

In the future we will work to better understand the other components of similarity for the concrete concepts, improve and refine the computational models of similarity for the Estonian language and address the same problem for different languages.

Reproducibility

The EstSimLex-999 set annotated with the human judge similarity scores, the code used to compute the results in this paper, the complete set of tables and the figures and most of the resources used in this paper can be referenced from the Github repository https://github.com/estsl/EstSimLex-999.

Acknowledgments

A number of persons have helped us with this study. We like to thank Aliis Aalmann and Kadri Muischnek for the aid in translation of the SimLex-999 set. We also like to thank Eleri Aedmaa for making available the word and sense embeddings and the anonymous annotators for rating the Estonian word pairs. This study was supported by the Estonian Ministry of Education and Research (IUT20-56).

References

Eduard Barbu, Heili Orav, and Kadri Vare. 2018. Topic interpretation using wordnet. In Kadri Muischnek and Kaili Müürisep, editors, *Baltic HLT*. IOS Press, volume 307 of *Frontiers in Artificial Intelligence and Applications*, pages 9–17.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Jrme Euzenat and Pavel Shvaiko. 2013. *Ontology Matching*. Springer Publishing Company, Incorporated, 2nd edition.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. *CoRR* abs/1506.05230. http://arxiv.org/abs/1506.05230.

J.R. Firth. 1961. *Papers in Linguistics 1934-1951: Repr.* Oxford University Press. https://books.google.ee/books?id=VxiWHAAACAAJ.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2016. Multiwibi: The multilingual wikipedia bitaxonomy project. *Artif. Intell.* 241:66–102.

Sebastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. 2015. *Semantic Similarity from Natural Language and Ontology Analysis*. Morgan & Claypool Publishers.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* pages 770–778.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* .

David H. Hubel and Torsten N. Wiesel. 1959. Receptive fields of single neurons in the cat's striate cortex. *Journal of Physiology* 148:574–591.

Heiki-Jaan Kaalep, Kadri Muischnek, Kristel Uiboaed, and Kaarel Veskis. 2010. The estonian reference corpus: Its composition and morphology-aware user interface. In *Proceedings of the 2010 Conference on Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*. IOS Press, Amsterdam, The Netherlands, The Netherlands, pages 143–146. http://dl.acm.org/citation.cfm?id=1860924.1860949.

Claudia Leacock and Martin Chodorow. 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*, volume 49, pages 265–283.

Ira Leviant and Roi Reichart. 2015. Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR* abs/1508.00106. http://arxiv.org/abs/1508.00106.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Curran Associates Inc., USA, NIPS'13, pages 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography* 3(4):235–244. ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf.

Siim Orasmaa, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. 2016. Estnltk - nlp toolkit for estonian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Berlin, Germany, pages 174–183. http://anthology.aclweb.org/W16-1620.

Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai. 2016. Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Berlin, Germany, pages 193–200. https://doi.org/10.18653/v1/W16-1622.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Beijing, China, pages 258–267. https://doi.org/10.18653/v1/K15-1026.

Stella Vosniadou and Andrew Ortony, editors. 1989. *Similarity and Analogical Reasoning*. Cambridge University Press, New York, NY, USA.

Max Wertheimer. 1938. Laws of organization in perceptual forms. In W. Ellis, editor, *A Source Book of Gestalt Psychology*, Routledge and Kegan Paul, London, pages 71–88.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '94, pages 133–138. https://doi.org/10.3115/981732.981751.