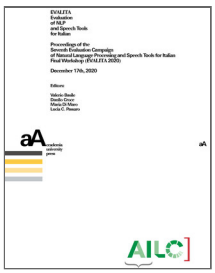

Valerio Basile, Danilo Croce, Maria Maro and Lucia C. Passaro (dir.)

EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020
Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop

Accademia University Press

ANDI @ CONcreTEXT: Predicting concreteness in context for English and Italian using distributional models and behavioural norms

Armand Stefan Rotaru

DOI: 10.4000/books.aaccademia.7465
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2020
Published on OpenEdition Books: 11 May 2021
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9791280136329

OpenEdition books

http://books.openedition.org

Electronic reference
ROTARU, Armand Stefan. *ANDI @ CONcreTEXT: Predicting concreteness in context for English and Italian using distributional models and behavioural norms* In: *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop* [online]. Torino: Accademia University Press, 2020 (generated 18 mai 2021). Available on the Internet: <http://books.openedition.org/aaccademia/7465>. ISBN: 9791280136329. DOI: https://doi.org/10.4000/books.aaccademia.7465.

View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by OpenEdition

ANDI @ CONcreTEXT: Predicting concreteness in context for English and Italian using distributional models and behavioural norms

Armand Stefan Rotaru
Independent researcher
armand.rotaru@gmail.com

Abstract

In this paper we describe our participation in the CONcreTEXT task of EVALITA 2020, which involved predicting subjective ratings of concreteness for words presented in context. Our approach, which ranked first in both the English and Italian subtasks, relies on a combination of context-dependent and context-independent distributional models, together with behavioural norms. We show that good results can be obtained for Italian, by first automatically translating the Italian stimuli into English, and then using existing resources for both Italian and English.

1 Introduction

In our everyday life we rarely encounter words in isolation. Instead, we typically process words as part of sentences or phrases, and these linguistic contexts shape our understanding of individual words. However, for various reasons, the overwhelming majority of behavioural norms that have been collected so far focus only on single words or word pairs (Johns et al., 2020).

Thus, the EVALITA 2020 (Basile et al., 2020) CONcreTEXT Task (Gregori et al., 2020) represents a timely and valuable contribution to the study of context-dependent semantics. The task asks competitors to predict subjective ratings of concreteness for words presented within sentences. As mentioned by the organizers, being able to automatically compute contextual concreteness ratings would have a several practical applications, such as identifying the use of figurative language, detecting words that might be difficult to understand for language learners, and allowing tighter control of contextual variables in psycholinguistic experiments.

In this paper we describe our computational models, based on pre-trained distributional models and behavioural norms, which ranked first in both the English and Italian tracks of the competition[1]. We find that the best performance can be obtained by employing a combination of transformer models, developed in the last 2 years. Moreover, for Italian, it is possible to reach good levels of performance by relying on both the original stimuli and their English translation, which allows access to resources for both languages.

1.1 General description

In order to predict concreteness in context, we use information derived from three type of sources, namely behavioural norms and distributional models, both context-independent (i.e., a model outputs the same vector representation for a given word, regardless of the context in which the word is encountered), and context-dependent (i.e., a model outputs a potentially different representations for a given word, as a function of the context in which the word is presented).

Firstly, we employ behavioural norms collected for a wide variety of psycholinguistic factors. Of particular interest to us are norms for concreteness (Brysbaert et al., 2014), semantic diversity (Hoffman et al., 2013), age of acquisition (Kuperman et al., 2012), emotional dimensions (i.e., valence, arousal, and dominance; Mohammad, 2018), and sensorimotor dimensions (i.e., modality strengths for the tactile, auditory, olfactory, gustatory, visual, and interoceptive modalities; interaction strengths for the mouth/throat, hand/arm, foot/leg, head excluding mouth/throat, and torso effectors; Lynott et al., 2019), as well as frequency and contextual diversity counts (Van Heuven et al., 2014).

[1] https://github.com/armandrotaru/TeamAndi-CONcreTEXT

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

We focus on these specific factors since they are meaningfully related to word concreteness (see the previous references).

Secondly, we employ context-independent distributional models, namely Skip-gram (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), and ConceptNet NumberBatch (Speer et al., 2017). Such models have been used in order to accurately predict a range of psycholinguistic variables, including concreteness (ρ = .88; Paetzold & Specia, 2016).

Thirdly, we employ context-dependent distributional models, namely BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2018), AlBERTo (Polignano et al., 2019), GPT-2 (Radford et al., 2019), Bart (Lewis et al., 2019), and ALBERT (Lan et al., 2020). Although they have become extremely popular after achieving human-level performance in various linguistic tasks (e.g., those in the GLUE benchmark; Wang et al., 2018), we are not aware of studies looking at whether such models can accurately predict (contextualized) subjective ratings. Nevertheless, since these models were specifically designed to process rich contextual information, they could be a valuable tool for predicting ratings of concreteness in context.

1.2 Predictors for English

We tested (combinations of) three groups of predictors. The first group was derived from large datasets of ratings for concreteness, semantic diversity, age of acquisition, emotional dimensions, and sensorimotor dimensions, as well as frequency and contextual diversity counts based on the SUBTLEX-UK and BNC corpora (see the references from the beginning of the previous section). In order to extend the coverage of the subjective ratings, we did not directly use them as predictors of concreteness in context. Instead, we relied on the Skip-gram, GloVe, and ConceptNet NumberBatch models, as a means of estimating the subjective ratings for more than 100,000 words, via linear regression. For the frequency and contextual diversity counts, we kept the original values, as they already have very good coverage. The intersection of the two datasets, which includes more than 70,000 words, served as the basis for our predictors of concreteness. More specifically, for each variable V (e.g., semantic diversity), we generated four predictors, namely $V(w)$, $V(c)$, $V(w) * V(c)$, and $abs(V(w) - V(c))$, where:

- $V(w)$ denotes the value of V corresponding to the word w (e.g., w = "offend"). If w is not present in our norms, we set $V(w)$ to the average value of V, computed over the entire norms;
- $V(c)$ denotes the value of V corresponding to the context c in which the word w is encountered (e.g., w = "offend", c = "Do not insult or ___ anyone ."). Computing this value involves calculating the average $V(c) = \frac{\sum_{i=1}^{N} V(c_i)}{N}$, where $V(c_i)$ is the value of V corresponding to the i-th context word, calculated as described previously, and N is the number of words that make up the context.

These predictors allowed us to include both the individual contributions of word w and its context c, as well as certain interactions between w and c.

The second group was derived from Skip-gram, GloVe, and ConceptNet NumberBatch embeddings, as well as from the concatenation of the three types of embeddings. The vocabulary of the four models is that described in the discussion above. Given the large number of dimensions involved (i.e., 300 + 300 + 300 + 900 = 1,800), we first extracted the top 20 principal components from each model (although comparable results can also be obtained by using a larger number of components). Then, for each variable V (e.g., PC_3 from the GloVe model) we generated four predictors, namely $V(w)$, $V(c)$, $V(w) * V(c)$, and $abs(V(w) - V(c))$, following the same procedure as in the previous discussion. In addition, based on (Frassinelli et al., 2017), for each distributional model we added four predictors based on a measure of neighbourhood density (i.e., the mean cosine similarity between a vector and its closest 20 vectors), using the same procedure as described above.

The third group was derived from the BERT, GPT-2, Bart, and ALBERT models. We used the standard (base) versions of each model (i.e., without task-specific fine-tuning), as described in the original papers, and obtained from the Hugging Face repository (https://huggingface.co/models).

Unlike for the previous two groups, the predictors consist only of a word's activations from the last hidden layer (i.e., for the GPT-2, Bart, and ALBERT models), or averaged from the last four hidden layers (i.e., for the BERT model).

Importantly, for each group of predictors we generated two sets of variables, based on two versions of the target words (i.e., the words rated by the participants). In the first set we used the uninflected form of the target words, taken from the TARGET column. In contrast, in the second set of we used the inflected form of the target words, taken from the words in the TEXT column located

at the positions specified in the INDEX column. More details can be found in Table 1.

For predicting ratings of concreteness in context, we employed ridge regression, with large values of the parameter lambda (i.e., strong regularization), after standardized all the variables.

1.3 Predictors for Italian

Our approach was similar to that for English, but with certain significant changes, as follows:

- for the first group of predictors, we began by automatically translating the Italian stimuli (i.e., the TARGET and TEXT columns) into English, using the MarianMT translation model (Junczys-Dowmunt et al., 2018). Next, for the translated stimuli we derived the predictors using the exact same procedure as in the case of English;
- for the second group of predictors, we employed Italian versions of the FastText and ConceptNet NumberBatch models), together with their concatenation. We derived the predictors based on the top 30 principal components for each model, rather than the top 20 principal components, as in the case of English (although comparable results can also be obtained by using a larger number of components);
- for the third group of predictors, we again employed the English translations and relied on the same models as for English, and also the RoBERTa model. For the BERT model, we only used the activations from the last hidden layer. We also added the AlBERTo model, but with the Italian stimuli.

As in the case for English, we generated two sets of predictors, using either the uninflected or inflected forms of the target words, together with their corresponding English translations. More details can be found in Table 1.

Once more, we employed ridge regression, with large values of the parameter lambda (i.e., strong regularization), after standardizing all the variables.

2 Results and discussion

The results for English and Italian are shown in Figures 1 and 2, respectively, for various sets of predictors and regularization strengths. Results are averaged over 1,000 rounds of 5-fold cross-validation, using only the training dataset.

For English, the results indicate that context-dependent models (Fig. 1c-d) outperform behavioural norms (Fig. 1a) and context-independent models (Fig. 1b). For the latter, even though we introduced contextual variables by averaging a given variable (e.g., concreteness) over the words that make up the context, it appears that this simple average does not properly capture contextual information and/or interactions between single word and contextual information. The addition the behavioural norms and/or context-independent models has a negligible effect on performance (Fig. 1e). In this respect, the excellent results for context-dependent models are likely due to several factors, such as the highly non-linear integration of contextual information, the use of attention mechanisms, and that of more sophisticated learning objectives (e.g., next sentence prediction).

Interestingly, predictors based on inflected targets consistently outperform those based on uninflected targets, especially for the context-dependent models. This shows that morphological information can be quite valuable. Also, even for the largest sets of predictors, consisting of more than 3,200 variables per 80 data points, the degree of regularization appears to matter very little, indicating surprisingly small levels of overfitting.

In the case of Italian, the findings are somewhat different from those for English. Performance is roughly 10% lower than that for English. This is expected, given that perfect translation from Italian to English is impossible, and that the majority of predictors depend on this translation. The gaps in performance between predictors for inflected vs uninflected targets (Fig. 2c-d), and between the various classes of predictors (Fig. 2a-e), are also smaller. Moreover, the performance of context-dependent models can be increased to a small degree by adding behavioural norms and/or context-independent models (Fig. 2f).

Our best models, as described in Figures 1 and 2, ranked first in both the English track ($\rho = .83$), and the Italian track ($\rho = .75$). The two correlations are smaller than those for the best models in the two figures, but this is likely to be an effect of distributional differences between the training set and the test set.

3 Conclusion

Our results suggest that a variety of approaches can be quite successfully employed in order to predict concreteness in context. The most effec-

tive predictors are those derived from context-dependent models (e.g., BERT), but relatively good results can be obtained also by using context-independent models (e.g., Skip-gram) and behavioural norms (e.g., ratings of semantic diversity).

Such an approach works very well for English, but less so for Italian, where the range of available predictors (i.e., pre-trained distributional models and large behavioural norms) is limited. One surprisingly effective solution to this problem is to simply translate the Italian stimuli into English, by relying on a neural machine translation system (e.g., MarianMT), and then make use of existing predictors for English. As an alternative to translating stimuli, it would be interesting to test whether comparable results can be obtained using multilingual versions of context-dependent models, such as BERT.

Acknowledgements

We would like to thank the anonymous reviewers, for their comments and suggestions, as well as the organizers of the competition, for their support.

Table 1. Type and number of predictors obtained from behavioural norms and distributional models. The same number of predictors are derived for both the inflected and uninflected versions of the target word. As predictors for the context-dependent models, we use the activations associated with the target, when presented in context (i.e., we do not have separate predictors for the target, context, and their potential interactions). More details regarding each set of predictors can be found in Subsections 2.2 and 2.3, as well as in Figures 1 and 2.

Predictors for English				
Source of predictors	# preds. V(w)	# preds. V(c)	# preds. V(w) * V(c)	# preds. abs(V(w) - V(c))
Behavioural norms (frequency, etc.)	20	20	20	20
Skip-gram (Google News – 100B)	21	21	21	21
GloVe (Common Crawl – 840B)	21	21	21	21
ConceptNet NumberBatch (ConceptNet + Skip-gram + GloVe)	21	21	21	21
Concatenation of Skip-gram, GloVe, and ConceptNet NumberBatch	21	21	21	21
ALBERT (last hidden layer)	768			
Bart (last hidden layer)	768			
BERT (last four hidden layers)	768			
GPT-2 (last hidden layer)	768			
Predictors for Italian				
Source of predictors	# preds. V(w)	# preds. V(c)	# preds. V(w) * V(c)	# preds. abs(V(w) - V(c))
Behavioural norms (frequency, etc.)	20	20	20	20
FastText (Common Crawl + Wikipedia)	31	31	31	31
ConceptNet NumberBatch (ConceptNet + Skip-gram + GloVe)	31	31	31	31
Concatenation of FastText and ConceptNet NumberBatch	31	31	31	31
ALBERT (last hidden layer)	768			
AlBERTo (last hidden layer)	768			
Bart (last hidden layer)	768			
BERT (last hidden layer)	768			
GPT-2 (last hidden layer)	768			
RoBERTa (last hidden layer)	768			

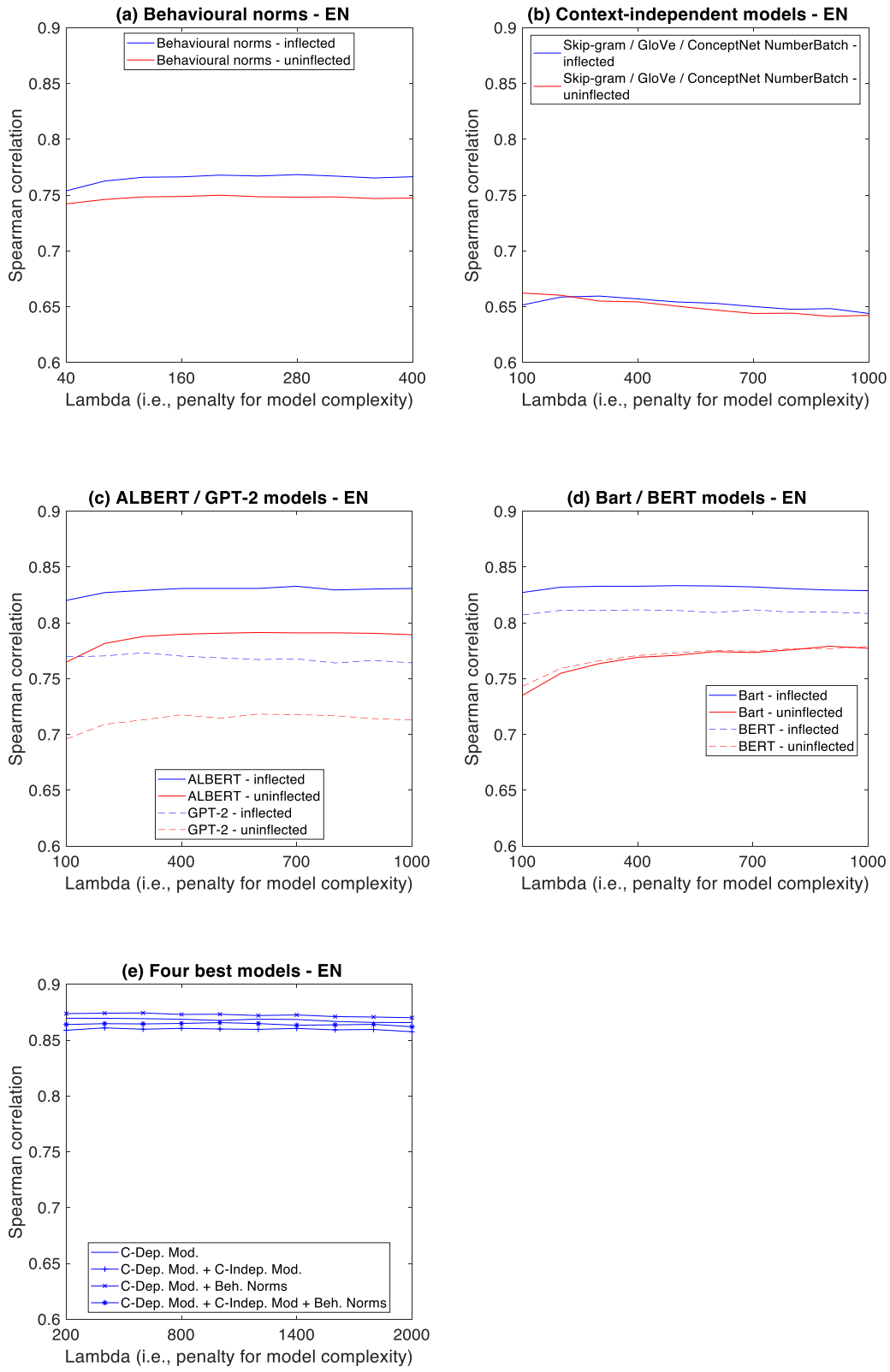

Figure 1: English: Spearman correlations between predicted and actual ratings, for various groups of predictors and regularization strengths (i.e., values of lambda). C-Dep. Mod.: the combination of the ALBERT, GPT-2, Bart, and BERT models; C-Indep. Mod.: the combination of the Skip-gram, GloVe, and ConceptNet NumberBatch models, their concatenation, and neighbourhood density measures; Beh. Norms: the predicted psycholinguistic ratings, together with frequency and contextual diversity counts. For the best four models, all predictors were derived from the inflected form of the target words. Our submission to the competition was based on C-Dep. Mod. + Beh. Norms (lambda = 500).

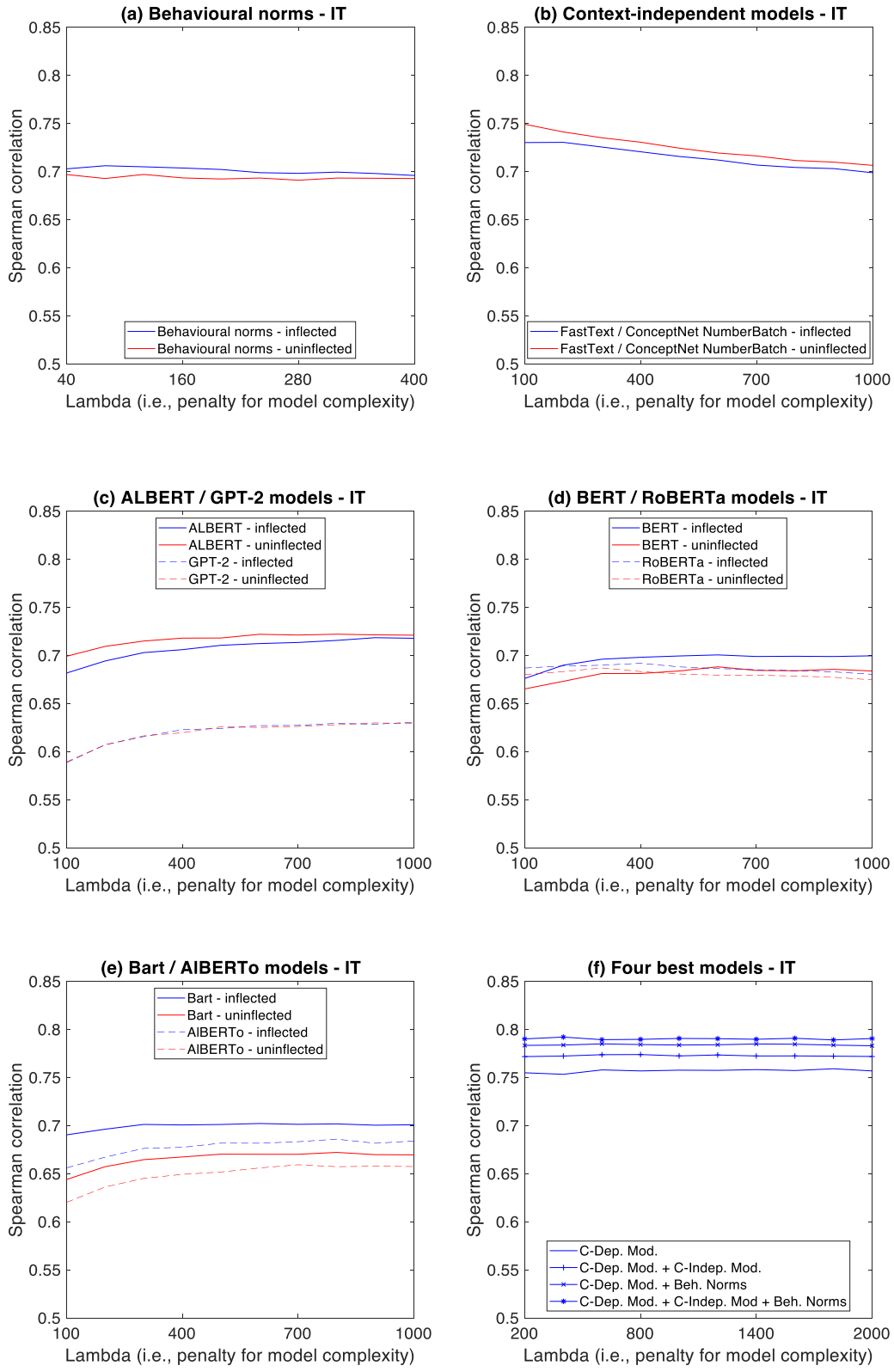

Figure 2. Italian: Spearman correlations between predicted and actual ratings, for various groups of predictors and regularization strengths (i.e., values of lambda). C-Dep. Mod.: the combination of the ALBERT, GPT-2, BERT, RoBERTa, Bart, and AlBERTo models; C-Indep. Mod.: the combination of the FastText and ConceptNet NumberBatch models, their concatenation, and neighbourhood density measures; Beh. Norms: the predicted psycholinguistic ratings, together with frequency and contextual diversity counts. For the best four models, all predictors were derived from the inflected form of the target words, except for the RoBERTa, FastText, and ConceptNet NumberBatch models (uninflected), and the behavioural norms (inflected and uninflected). Our submission to the competition was based on C-Dep. Mod. + C-Indep. Mod. + Beh. Norms (lambda = 500).

References

Basile, V., Croce, D., Di Maro, M., Passaro, L.C., 2020. EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Basile, V., Croce, D., Di Maro, M., Passaro, L.C. (Eds.), Proceedings of 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020). CEUR.org, Online.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the NAACL-HLT* (pp. 4171-4186). Stroudsburg, PA: ACL.

Frassinelli, D., Naumann, D., Utt, J., & im Walde, S. S. (2017). Contextual characteristics of concrete and abstract words. In C. Gardent & C. Retoré (Eds.), *Proceedings of the IWCS* (pp. 1-7). Stroudsburg, PA: ACL.

Gregori, L., Montefinese, M., Radicioni, D. P., Ravelli, A. A., & Varvara, R. (2020). CONcreTEXT @ Evalita2020: the Concreteness in Context Task. In V. Basile, D. Croce, M. Di Maro, & L. C. Passaro (Eds.), *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*. Online: CEUR.org.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718-730.

Johns, B. T., Jamieson, R. K., & Jones, M. N. (2020). The continued importance of theory: Lessons from big data approaches to language and cognition. In S. E. Woo, R. Proctor, & L. Tay (Eds.), *Big data methods for psychological research: New horizons and challenges* (pp. 277-295). Washington, DC: APA.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A., & Birch, A. (2018). Marian: Fast neural machine translation in C++. In F. Liu & T. Solorio (Eds.), *Proceedings of the ACL - System Demonstrations* (pp. 116-121). Stroudsburg, PA: ACL.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the ICLR* (pp. 1-17).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the ACL* (pp. 7871-7880). Stroudsburg, PA: ACL.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis. M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint:1907.11692*.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*, 1-21.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In J. Bengio & Y. LeCun (Eds.), *Proceedings of the Workshop at the ICLR* (pp. 1-12).

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the ACL - Long Papers* (pp. 174-184). Stroudsburg, PA: ACL.

Paetzold, G., & Specia, L. (2016). Inferring psycholinguistic properties of words. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the NAACL-HLT* (pp. 435-440). Stroudsburg, PA: ACL.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the EMNLP* (pp. 1532-1543). Stroudsburg, PA: ACL.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019). AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In R. Bernardi, R. Navigli, & G. Semeraro (Eds.), *Proceedings of CLiC-it*. Aachen, Germany: CEUR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In S. P. Singh & S. Markovitch (Eds.), *Proceedings of the AAAI* (pp. 4444-4451). Palo Alto, CA: AAAI Press.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the EMNLP Workshop BlackboxNLP* (pp. 353-355). Stroudsburg, PA: ACL.