CAPISCO @ CONcreTEXT 2020: (Un)supervised Systems to Contextualize Concreteness with Norming Data

Alessandro Bondielli[1,2] and **Gianluca E. Lebani** [3] and **Lucia C. Passaro**[2] and **Alessandro Lenci**[2]

[1] Dipartimento di Ingegneria dell'Informazione, Università degli studi di Firenze
[2] CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa
[3] Dipartimento di Studi Linguistici e Culturali Comparati, Università Ca' Foscari Venezia

alessandro.bondielli@unifi.it
gianluca.lebani@unive.it
lucia.passaro@fileli.unipi.it
alessandro.lenci@unipi.it

Abstract

English. This paper describes several approaches to the automatic rating of the concreteness of concepts in context, to approach the EVALITA 2020 "CONcreTEXT" task. Our systems focus on the interplay between words and their surrounding context by (i) exploiting annotated resources, (ii) using BERT masking to find potential substitutes of the target in specific contexts and measuring their average similarity with concrete and abstract centroids, and (iii) automatically generating labelled datasets to fine tune transformer models for regression. All the approaches have been tested both on English and Italian data. Both the best systems for each language ranked second in the task.

1 Introduction

The characterization of the conceptual concreteness of a word in context is a task that requires a level of analysis that goes well beyond the identification of the properties of the referent (or denotation) of the target word. The overall linguistic context should be taken into consideration as well, along with its interaction with the target word. Even addressed in the most simplistic way, i.e. ignoring the context and focusing solely on the target word in isolation, it is a daunting task in which the machine is asked to draw inferences on a level of semantic representation that the speaker builds by integrating experiential and linguistic information (Vigliocco et al., 2009). Moreover, figurative uses of words (e.g., metaphors) determine important shifts in their concreteness values. For example, the word *head* in the sentence *Take your safety pins and attach one card to the head of your bed* can be considered as highly concrete, as it describes a physical object. Conversely, the same word in the sentence *The pope is also head of the world's smallest sovereign state, The Vatican* has a more abstract meaning, denoting the title of a person. Similarly the verb *fly* is more concrete in the sentence *The plane flies in the sky* than in the metaphorical sentence *Time flies*.

The context-sensitive nature of word concreteness is one of the key elements that make its identification very interesting and complex from a Natural Language Processing (NLP) perspective (Naumann et al., 2018). Unfortunately, to the best of our knowledge only a handful of scholars have addressed this topic. Notable mentions are Hill et al. (2013), and Hill and Korhonen (2014).

As it is common for other NLP and NLP-related tasks and topics, an invaluable source of knowledge that can be used both to train models and to gain some insights on the nuances of the problem itself can be found in the psycho-linguistic tradition, and especially in those normative studies built to analyze collections of human-elicited concreteness judgements (Brysbaert et al., 2013; Montefinese et al., 2013; Della Rosa et al., 2010). Most of these works, however, share the common limitation of ignoring the polysemic nature of words and the effect of context on their concreteness (Reijnierse et al., 2019). As an NLP task, the automatic estimation of the degree of concreteness carried by a given word in a given linguistic context can play a part in well-known and long-standing NLP issues such as word sense disam-

biguation (Agirre and Edmonds, 2007) and figurative language interpretation (Veale et al., 2016). All such tasks require a deep understanding of the linguistic context and are quite hard to model with traditional NLP models. Moreover, the fortune of language models specifically focused on modelling the meaning of words in context, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), demonstrates how meaning construction is an appealing topic for the whole NLP community.

The CONcreTEXT task (Gregori et al., 2020) of EVALITA 2020 (Basile et al., 2020) focuses on modelling the concreteness of concepts in context. Given a sentence and a target word, the goal is to predict the word concreteness on a scale from 1 (fully abstract) to 7 (fully concrete). Results are evaluated by estimating their Spearman correlation with the (average of the) human-generated ratings. For the task, two trial datasets were made available, one for English and the other for Italian. Each trial dataset contains 100 sentences, two for each of the 50 target words.

In order to address this task, we propose three families of distributional semantic methods relying on several existing concreteness norms. Our general approach revolves around the idea that taking into account both the context and the target word, as well as words that play a similar role in the same context, may help us in overcoming limitations due to scarce training data, and may prove beneficial for predicting more accurate ratings.

The paper is organized as follows: Section 2 describes the proposed approach based on both supervised and unsupervised methods. Section 3 presents the results, which are discussed in Section 4. Finally, Section 5 draws some conclusions.

2 Methods

We propose three different "CAPISCO" (for *CA' Foscari and PISa COncretext project*) approaches for predicting the concreteness of a word in a given context of occurrence. Each method exploits the assumption that the concreteness of a word is influenced by its surrounding context. We explore both unsupervised and supervised techniques. In fact, two such approaches are unsupervised, and exploit either pre-trained word embeddings, or pre-trained transformer language models, while the third method is supervised:

NON-CAPISCO – the concreteness of the target word is modelled as a function of its concreteness value in isolation and of the average concreteness of its surrounding context.

CAPISCO-CENTROIDS – the concreteness of the target word is estimated as a function of the concreteness values of its closer synonyms according to a pre-trained transformer language model. Crucially, the concreteness ratings of the target synonyms are estimated by computing their distance from two reference points in the distributional space corresponding to the centroids of the highly concrete and highly abstract terms.

CAPISCO-TRANSFORMER – a supervised regressor is trained to predict concreteness ratings. Specifically, we fine-tune a transformer model to predict the target concreteness of the sentence, exploiting the available dataset augmented with new data automatically generated from several different norms of concreteness.

2.1 NON-CAPISCO

The NON-CAPISCO system is rather simple, both conceptually and implementation-wise. It is based on a minor change in the baseline proposed by the task organizers.

The task baseline is computed by averaging over the concreteness ratings of all the words in the sentence. Ratings are obtained from the norms by Montefinese et al. (2013) for Italian, and from those by Brysbaert et al. (2013) for English. Words missing from these resources are replaced by their closest neighbor among those for which human ratings are available. Closest neighbors are identified using fastText (Grave et al., 2018). On the trial dataset, our implementation of the baseline obtained a Spearman correlation score of 0.47 for Italian and 0.57 for English.

Crucially, this baseline takes into account the concreteness rating of the target word, but it has the same weight as all the other words in the sentence on the final prediction. On the other hand, we noticed that a simple method based solely on the concreteness score of the target word achieves a performance of 0.69 for Italian and 0.69 for English, much higher than that of the task baseline. This led us to surmise that, at least in the task dataset, the concreteness of the word in context is strongly affected by its value in isolation.

The NON-CAPISCO method gives more weight to the target word, by multiplying its concreteness rating for the mean concreteness of the whole sen-

tence. On the trial dataset this combined score obtained a Spearman correlation of 0.73 for Italian and 0.73 for English.

2.2 CAPISCO-CENTROIDS

The CAPISCO-CENTROIDS approach is based on the assumption that semantically similar words are expected to be similarly rated for concreteness and that, conversely, words associated with highly different concreteness scores should be placed far away from each other in semantic space. This assumption is driven by the fact that concrete (or abstract) senses are typically found in co-occurrence with other concrete (or abstract) ones (Frassinelli et al., 2017). Thus, semantically similar words, i.e. that typically occur in the same context, are expected to have similar concreteness as well.

The first step of this method consisted in the building of two reference vectors: one representing the prototypical abstract concept; the other representing the prototypical concrete concept. To this end, we first identified highly concrete and highly abstract terms from two available resources: the Brysbaert et al. (2013) norms for English and the Della Rosa et al. (2010) norms for Italian. The latter has been preferred to more comprehensive alternatives, like the Montefinese et al. (2013) norms, due to its covering of a significant set of highly polarized words.

For each resource, the clusters of most concrete and abstract words were identified by fitting a mixture-of-Gaussian model on the human judgments, and choosing the most distant clusters. We used the expectation-maximization algorithm available in scikit-learn.[1] To set the number of clusters and type of covariance, we chose the pair that minimized the Bayesian information criterion. After identifying the groups of most polarized words in our reference norm, we used English and Italian pre-trained word embeddings from fastText (Grave et al., 2018) to identify their respective centroids in the vector space, by simply averaging the embeddings of highly concrete and highly abstract words. In the case of the English vector space, the dimensionality was left to the default value of 300. In the case of the Italian space, the dimensionality was further reduced to 100, as we saw an increase in performances, which instead was not the case for English.

However, predicting the concreteness of a target word solely based on its proximity with the centroids could be biased by its semantic relatedness with the words used for building the centroids. To smooth this bias, the final score for a given target word was calculated as the average of the similarities of its potential lexical substitutes. BERT was used to identify the substitutes of each target word in context. Operationally, we masked the target word in each sentence, and asked the model to predict the 50 most likely words that may fill the masked token, which is likely to include the target itself. After several experiments, we chose 50 words as they gave us the best overall results. We can argue that it is probably the best trade-off between number of neighbors and their actual similarity with the target word. We used the `bert-base-uncased` model for English, and the `bert-base-italian-xxl-uncased` model for Italian. Table 1 reports some potential substitutes of the target word in the sentence.

TARGET	MASKED SENT.	FILLERS
lawsuit	In a typical [MASK] , the defendant frequently brings a motion [...].	case trial proceeding
love	Give your friends [MASK] , positivity , and compliments .	attention kindness respect

Table 1: Prediction of fillers in context with BERT.

To avoid noise due to the fact that sometimes BERT predicts a token with a different syntactic role, all the fillers with a different Part-of-Speech (PoS) tag than that of the target word were filtered out. To this end, we PoS-tagged all the sentences produced by replacing the target word and kept only those with the same PoS sequence of the original sentence. This way, we obtained, for each target word, a list of lexical substitutes in a particular context. Each substitute was assigned a concreteness score based on its proximity to the two prototypical vectors. More specifically, we computed the concreteness of a word as the absolute value of the difference between its cosine with the concrete centroid and its cosine with the abstract one normalized on a 1-7 scale. Finally, each target word was assigned with a concreteness value obtained by averaging the concreteness of its substitutes.

2.3 CAPISCO-TRANSFORMER

The CAPISCO-TRANSFORMER system addresses the problem from a supervised perspective. The system is based on the BERT Transformer archi-

[1] `https://scikit-learn.org/`

tecture (Devlin et al., 2019). BERT and the other Transformer allow for transfer learning in NLP tasks, by means of unsupervised pre-training followed by supervised fine-tuning for downstream tasks. Such models have obtained state-of-the-art results in most NLP supervised and unsupervised tasks (Devlin et al., 2019). We used a BERT pre-trained model and fine-tuned it on the concreteness rating task. Given the very small size of the trial dataset provided for the task, we tried to improve generalization capabilities by dynamically generating additional training data to feed the model. To this end, we used two different approaches.

On the one hand, we generated potential substitutes of the target word with the same techniques used in Section 2.2. In this case, we generated three sentences containing as target word the three most likely lexical substitutes of the original one. Such new target words were assigned the same concreteness rating of the original one, modified by a small random value in the range [-0.2,0.2], to avoid repetition of target values for the training set derived from the gold data.

On the other hand, we extended the dataset with new sentences which were assigned the concreteness scores found in the concreteness norm. For English, we extracted from the BNC corpus (The British National Corpus, 2007) all the sentences containing words rated in the Brysbaert et al. (2013) norms. For Italian language, we extracted from La Repubblica corpus (Baroni et al., 2004) all the sentences containing words rated in the Montefinese et al. (2013) or in the Della Rosa et al. (2010) norms. As we are interested in mostly unambiguous target words with different concreteness ratings, we chose to select, for each considered norm, only words with a low standard deviation that are in a specific range of values for concreteness. Therefore, we obtained three sets of very concrete, very abstract and mildly concrete words. Thresholds were manually set for each resource in order to address their different distribution and scales in terms of concreteness ratings. Once sentences containing such target words were collected, we sampled three random sentences for each target and we assigned each sentence the concreteness rating of its target word in the norm. We obtained 8,813 training sentences for English and 3,467 for Italian. The Italian training set is smaller as the Italian resources contain fewer words.

The whole extended dataset is then used to fine-tune the BERT model to predict the concreteness rating assigned to the whole sentence by means of regression. Operationally, we use the implementation of BERT provided in the Huggingface library.[2] For the English model, initial weights are taken from `bert-base-uncased`, while for Italian we used initial weights from `bert-base-italian-xxl-uncased`. Both pre-trained models are available within the Transformer library. We trained each model for 2 epochs, with a batch size of 8 and the learning rate set to 2e-5, on a machine equipped with a Titan Xp GPU. At inference time, we simply feed the fine-tuned model with test sentences and ask it to directly predict the concreteness rating.

3 Results

We proposed three different approaches for the estimation of concreteness. The performances obtained for each model for the Italian language and for the English language are presented respectively in Tables 2 and 3. Given the absence of a training set, we decided to give more emphasis to the unsupervised method (NON-CAPISCO) based on the concreteness of target words and of the surrounding context. It is clear that the results of this method are highly influenced by the annotated resources exploited to infer the concreteness. The results revealed that while for English such approach was quite effective, for Italian it is not, probably due to the smaller dimension and quality of the resources taken into consideration. In fact, if we look at the ranking of our models in the two languages, the results are reversed. On the one hand, the best CAPISCO approach for English is the NON-CAPISCO system, in which concreteness ratings are obtained from Brysbaert et al. (2013). Such resource counts ratings for about 40 thousand of English lemmas that have been annotated for several variables. On the other hand, the Italian resources (Della Rosa et al., 2010; Montefinese et al., 2013) are orders of magnitude smaller than English ones thus causing a big drop in performances of the proposed approach. This issue will be discussed in detail in Section 4.

4 Discussion

In light of the reported results, several interesting observations can be made. For both languages, our best-performing model ranked sec-

[2] `https://huggingface.co`

RANK	SYSTEM	SPEARMAN
1	****	0.749
2	CAPISCO-TRANSFORMER-IT	0.625
3	CAPISCO-CENTROIDS-IT	0.615
4	NON-CAPISCO-IT	0.557
5	Baseline_2	0.534
6	Baseline_1	0.346

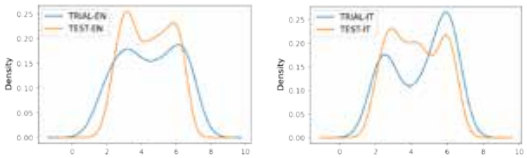
Table 2: CAPISCO performances for the Italian.

RANK	SYSTEM	SPEARMAN
1	****	0.833
2	NON-CAPISCO-EN	0.785
3	****	0.663
4	****	0.651
5	Baseline_2	0.554
6	****	0.542
7	CAPISCO-CENTROIDS-EN	0.542
8	****	0.541
9	CAPISCO-TRANSFORMER-EN	0.504
10	Baseline_1	0.383
11	****	-0.013
12	****	-0.124
13	****	-0.127

Table 3: CAPISCO performances for the English.

ond overall. However, we can notice how neither numerical results nor the ranking of the system are consistent across languages. For English, the best performing system is NON-CAPISCO. The system strongly outperforms both baselines and the other two methods. We must also note that both CAPISCO-CENTROIDS and CAPISCO-TRANSFORMER perform worse than one of the two baselines. On the other hand, for Italian, CAPISCO-TRANSFORMER performed best, closely followed by CAPISCO-CENTROIDS. Both outperform the NON-CAPISCO approach, and all three systems perform better than the baselines. This discrepancy may be due to several key aspects concerning both the resources used as well as some crucial differences among trial and test samples of the dataset.

We can identify several key differences among English and Italian resources that may justify such drastically different performances. While for English a comprehensive resource with 40,000 words is available, both resources for Italian are orders of magnitude smaller. In addition to this, especially for ratings contained in Montefinese et al. (2013), the distribution is unbalanced towards mid-range and high values of concreteness, while ratings for Brysbaert et al. (2013) are more evenly distributed across the spectrum. For the NON-CAPISCO system, this may lead to poor performances since for the system is more difficult to predict higher values for the Italian dataset. While predictions for the English model closely follow the distribution of ratings in the test set, predictions for Italian are unbalanced towards lower values.



(a) (b)

Figure 1: Distribution of ratings for trial and test sets, for (a) English and (b) Italian.

On the contrary, for the CAPISCO-CENTROIDS system, this has the opposite effect. In fact, given that it is more difficult to isolate extremely abstract and extremely concrete terms, centroids built from Italian resources are closer one another, and thus prediction based on the difference between distances to the centroids almost always fall in the middle of the range, while for English the same approach has the effect of yielding results that are mostly close to the lower-end of the spectrum. This, in turn, has the effect of seemingly improving performances for Italian, because too high and too low prediction balance each other, while errors for English are more pronounced.

Finally, for the CAPISCO-TRANSFORMER system, it may be possible that the fact that English norms contains more high frequency words, may hinder the generalization capabilities of the model. In fact, if such words are found in very different sentences, all such sentences are assigned very similar concreteness scores and the predictions are biased towards certain values for many different sentences. Therefore, the distribution of predictions follow the same tripartite distribution of the sampled words in terms of concreteness.

Finally, we must point out that the distribution of ratings in the trial and test set are rather different, as shown in Figure 1. This may have hindered our judgment on the quality of all proposed systems, both unsupervised and supervised.

5 Conclusions and Future works

The models proposed are based on both supervised and unsupervised approaches. The choice was motivated by the fact that the trial dataset proposed for the task is too small to effectively train supervised learning models on it. The key assumption

that drove the development is that the concreteness of a word is influenced by its surrounding context, as claimed by the task organizers as well. The best CAPISCO systems for both Italian and English ranked second in the CONcreTEXT task despite the fact that results differ a lot in terms of absolute performances and used method. For Italian, the best CAPISCO system is based on Transformers and reaches a Spearman correlation of 0.625 with gold data. The best CAPISCO model for English, on the contrary, is unsupervised and reaches a Spearman correlation with gold data of 0.785.

In the future, we plan to perform some additional hyper-parameter tuning on the models. Moreover, we would like to test this approach in similar tasks (e.g. predicting abstractness). We are confident that by exploiting the dynamic selection of training data in addition to an annotated dataset such as the test dataset provided by the task organizers would improve the results of our systems, and in particular of the transformers-based one.

Acknowledgments

We gratefully acknowledge the support of the NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Eneko Agirre and Philip Edmonds. 2007. *Word Sense Disambiguation: Algorithms and applications*. Springer Science & Business Media.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian. *Proc. of LREC 2004*.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proc. of EVALITA 2020*, Online. CEUR.org.

Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behav. Res. Methods*, 46:904–911.

Pasquale Della Rosa, Eleonora Catricalà, Gabriella Vigliocco, and Stefano Cappa. 2010. Beyond the abstract-concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 italian words. *Behav. Res. Methods*, 42:1042–1048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT 2019*, pages 4171–4186.

Diego Frassinelli, Daniela Naumann, J. Utt, and Sabine Schulte im Walde. 2017. Contextual characteristics of concrete and abstract words. In *Proc. of IWCS 2017*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proc. of LREC 2018*.

Lorenzo Gregori, Maria Montefinese, Daniele P. Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. CONCRETEXT @ EVALITA2020: the Concreteness in Context Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proc. of EVALITA 2020)*, Online. CEUR.org.

Felix Hill and Anna Korhonen. 2014. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proc. of EMLP 2014*, pages 255–265.

Felix Hill, Douwe Kiela, and Anna Korhonen. 2013. Concreteness and corpora: A theoretical and practical study. In *Proc. of CMCL 2013*, pages 75–83.

Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2013. The adaptation of the affective norms for english words (anew) for italian. *Behav. Res. Methods*, 46:887–903, 10.

Daniela Naumann, Diego Frassinelli, and Sabine Schulte im Walde. 2018. Quantitative semantic variation in the contexts of concrete and abstract words. In *Proc. of STARSEM 2018*, pages 76–85, New Orleans, Louisiana, June. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL 2018*, page 2227–2237.

W. Gudrun Reijnierse, Christian Burgers, Marianna Bolognesi, and Tina Krennmayr. 2019. How polysemy affects concreteness ratings: The case of metaphor. *Cognitive Science*, 43(8):e12779.

The British National Corpus. 2007. version 3 (BNC XML Edition).

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: A Computational Perspective*. Morgan & Claypool.

Gabriella Vigliocco, Lotte Meteyard, Mark Andrews, and Stavroula Kousta. 2009. Toward a theory of semantic representation. *Language and Cognition*, 1(2):219–247.