A Comparison of Representation Models in a Non-Conventional Semantic Similarity Scenario

Andrea Amelio Ravelli
University of Florence
andreaamelio.ravelli@unifi.it

Oier Lopez de Lacalle and **Eneko Agirre**
University of the Basque Country
e.agirre@ehu.eus
oier.lopezdelacalle@ehu.eus

Abstract

Representation models have shown very promising results in solving semantic similarity problems. Normally, their performances are benchmarked on well-tailored experimental settings, but what happens with unusual data? In this paper, we present a comparison between popular representation models tested in a non-conventional scenario: assessing action reference similarity between sentences from different domains. The action reference problem is not a trivial task, given that verbs are generally ambiguous and complex to treat in NLP. We set four variants of the same tests to check if different pre-processing may improve models performances. We also compared our results with those obtained in a common benchmark dataset for a similar task.[1]

1 Introduction

Verbs are the standard linguistic tool that humans use to refer to actions, and action verbs are very frequent in spoken language (~50% of total verbs occurrences) (Moneglia and Panunzi, 2007). These verbs are generally ambiguous and complex to treat in NLP tasks, because the relation between verbs and action concepts is not one-to-one: e.g. (a) *pushing a button* is cognitively separated from (b) *pushing a table to the corner*; action (a) can also be predicated through *press*, while *move* can be used for (b) and not vice-versa (Moneglia, 2014). These represent two different *pragmatic actions*, despite of the verb used to describe it, and all the possible objects that can undergo the action. Another example could be the ambiguity behind a sentence like *John pushes the bottle*: is the agent applying a continuous and controlled force to move the object from position A to position B, or is he carelessly shoving an object away from its location? These are just two of the possible interpretation of this sentence *as is*, without any other lexical information or pragmatic reference.

Given these premises, it is clear that the task of automatically classifying sentences referring to actions in a fine-grained way (e.g. *push/move* vs. *push/press*) is not trivial at all, and even humans may need extra information (e.g. images, videos) to precisely identify the exact action. One way could be to consider action reference similarity as a Semantic Textual Similarity (STS) problem (Agirre et al., 2012), assessing that lexical semantic information encodes, at a certain level, the action those words are referring to. The simplest way is to make use of pre-computed word embeddings, which are ready to use for computing similarity between words, sentences and documents. Various models have been presented in the past years that make use of well-known static word embeddings, like word2vec, GloVe and FastText (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017). Recently, the best STS models rely on representations obtained from contextual embeddings, such as ELMO, BERT and XLNet (Peters et al., 2018; Devlin et al., 2018; Yang et al., 2019).

In this paper, we are testing the effectiveness of representation models in a non-conventional scenario, in which we do not have labeled data to train STS systems. Normally, STS is performed on sentence pairs that, on one hand, can have very close or distinct meaning, i.e. the assertion of similarity is easy to formulate; on the other hand, all sentences derive from the same domain, thus they share some syntactic regularities and vocabulary. In our scenario, we are computing STS between textual data from two different resources, IMAGACT and LSMDC16 (described respectively in

5.1 and 5.2), in which the language used is highly different: from the first, synthetic and short captions; from the latter, audio descriptions. The objective is to benchmark word embedding models in the task of estimating the action concept expressed by a sentence.

2 Related Works

Word embeddings are abstract representations of words in the form of dense vectors, specifically tailored to encode semantic information. They represent an example of the so called transfer learning, as the vectors are built to minimize certain objective function (i.e., guessing the next word in a sentence), but successfully applied on different unrelated tasks, such as searching for words that are semantically related. In fact, embeddings are typically tested on semantic similarity/relatedness datasets, where a comparison of the vectors of two words is meant to mimic a human score that assesses the grade of semantic similarity between them.

The success of word embeddings on similarity tasks has motivated methods to learn representations of longer pieces of text such as sentences (Pagliardini et al., 2017), as representing their meaning is a fundamental step on any task requiring some level of text understanding. However, sentence representation is a challenging task that has to consider aspects such as compositionality, phrase similarity, negation, etc. The Semantic Textual Similarity (STS) task (Cer et al., 2017) aims at extending traditional semantic similarity/relatedness measures between pair of words in isolation to full sentences, and is a natural dataset to evaluate sentence representations. Through a set of campaigns, STS has distributed set of manually annotated datasets where annotators measure the similarity among sentences with a score that ranges between 0 (no similarity) to 5 (full equivalence).

In the recent years, evaluation campaigns that agglutinate many semantic tasks have been set up, with the objective to measure the performance of many natural language understanding systems. The most well-known benchmarks are SentEval[2](Conneau and Kiela, 2018) and GLUE[3] (Wang et al., 2019). They share many of existing tasks and datasets, such as sentence similarity.

3 Problem Formulation

We cast the problem as a fine-grained action concept classification for verbs in LSMDC16 captions (e.g. *push* as *move* vs *push* as *press*, see Figure 1). Given a caption and the target verb from LSMDC16, our aim is to detect the most similar caption in IMAGACT that describe the action. The inputs to our model are the target caption and an inventory of captions that categorize the possible action concepts of the target verb. The model ranks the captions in the inventory according to the textual similarity with the target caption, and, similar to a kNN classifier, the model assigns the action label of *k* most similar captions.

4 Representation Models

In this section we describe the pretrained embeddings used to represent the contexts. Once we get the representation of each caption, the final similarity is computed based on cosine of the two representation vectors.

4.1 One-hot Encoding

This is the most basic textual representation, in which text is represented as binary vector indicating the words occurring in the context (Manning et al., 2008). This way of representing text creates long and sparse vectors, but it has been successfully used in many NLP tasks.

4.2 GloVe

The Global Vector model (GloVe)[4] (Pennington et al., 2014) is a log-linear model trained to encode semantic relationships between words as vector offsets in the learned vector space, combining global matrix factorization and local context window methods.

Since GloVe is a word-level vector model, we compute the mean of the vectors of all words composing the sentence, in order to obtain the sentence-level representation. The pre-trained model from GloVe considered in this paper is the 6B-300d, counting a vocabulary of 400k words with 300 dimensions vectors and trained on a dataset of 6 billion tokens.

[2] https://github.com/facebookresearch/SentEval

[3] https://gluebenchmark.com/

[4] https://nlp.stanford.edu/projects/glove/

4.3 BERT

The Bidirectional Encoder Representations from Transformer (BERT)[5] (Devlin et al., 2018) implements a novel methodology based on the so called *masked language model*, which randomly masks some of the tokens from the input, and predicts the original vocabulary id of the masked word based only on its context.

Similarly with GloVe, we extract the token embeddings of the last layer, and compute the mean vector to obtain the sentence-level representation. The BERT model used in our test is the BERT-Large Uncased (24-layer, 1024-hidden, 16-heads, 340M parameters).

4.4 USE

The Universal Sentence Encoder (USE) (Cer et al., 2018) is a model for encoding sentences into embedding vectors, specifically designed for transfer learning in NLP. Based on a deep averaging network encoder, the model is trained for a variety text length, such as sentences, phrases or short paragraphs, and in a variety of semantic task including the STS. The encoder returns the corresponding vector of the sentence, and we compute similarity using cosine formula.

5 Datasets

In this section, we briefly introduce the resources used to collect sentence pairs for our similarity test. Figure 1 shows some examples of data, aligned by action concepts.

5.1 IMAGACT

IMAGACT[6] (Moneglia et al., 2014) is a multilingual and multimodal ontology of action that provides a video-based translation and disambiguation framework for action verbs. The resource is built on an ontology containing a fine-grained categorization of action concepts (*ac*s), each represented by one or more visual prototypes in the form of recorded videos and 3D animations. IMAGACT currently contains 1,010 scenes, which encompass the actions most commonly referred to in everyday language usage.

Verbs from different languages are linked to *ac*s, on the basis of competence-based annotation from mother tongue informants. All the verbs that productively predicates the action depicted in an *ac* video are in *local equivalence* relation (Panunzi et al., 2018b), i.e the property that different verbs (even with different meanings) can refer to the same action concept. Moreover, each *ac* is linked to a short *synthetic* caption (e.g. *John pushes the button*) for each locally equivalent verb in every language. These captions are formally defined, thus they only contain the minimum arguments needed to express an action.

We exploited IMAGACT conceptualization due to its *action-centric* approach. In fact, compared to other linguistic resources, e.g. WordNet (Fellbaum, 1998), BabelNet (Navigli and Ponzetto, 2012), VerbNet (Schuler, 2006), IMAGACT focuses on actions and represents them as visual concepts. Even if IMAGACT is a smaller resource, its action conceptualization is more fine-grained. Other resources have more broad scopes, and for this reason senses referred to actions are often vague and overlapping (Panunzi et al., 2018a), i.e. all possible actions can be gathered under one synset. For instance, if we look at the senses of *push* in Wordnet, we find that only 4 out of 10 synsets refer to concrete actions, and some of the glosses are not really exhaustive and can be applied to a wide set of different actions:

- push, force (move with force);
- push (press against forcefully without moving);
- push (move strenuously and with effort);
- press, push (make strenuous pushing movements during birth to expel the baby).

In such framework of categorization, all possible actions referred by *push* can be gathered under the first synset, except from those specifically described by the other three.

For the experiments proposed in this paper, only the English captions have been used, in order to test our method in a monolingual scenario.

5.2 LSMDC16

The Large Scale Movie Description Challenge Dataset[7] (LSMDC16) (Rohrbach et al., 2017) consists in a parallel corpus of 128,118 sentences obtained from audio descriptions for visually impaired people and scripts, aligned to video clips

[5] `https://github.com/google-research/bert`

[6] `http://www.imagact.it`

[7] `https://sites.google.com/site/describingmovies/home`

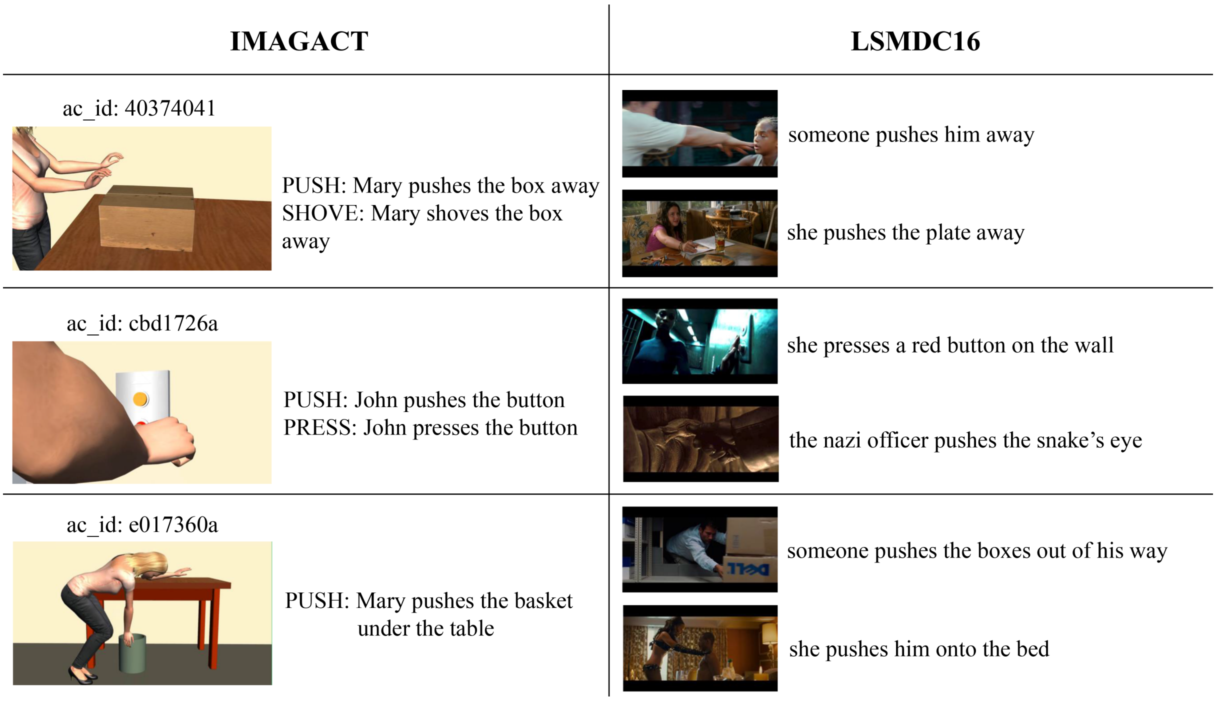
IMAGACT
LSMDC16
ac_id: 40374041
PUSH: Mary pushes the box away
SHOVE: Mary shoves the box away
someone pushes him away
she pushes the plate away
ac_id: cbd1726a
PUSH: John pushes the button
PRESS: John presses the button
she presses a red button on the wall
the nazi officer pushes the snake's eye
ac_id: e017360a
PUSH: Mary pushes the basket under the table
someone pushes the boxes out of his way
she pushes him onto the bed

Figure 1: An example of aligned representation of action concepts in the two resources. On the left, action concepts with prototype videos and captions for all applicable verbs in IMAGACT; on the right, the video-caption pairs in LSMDC16, classified according to the depicted and described action.

from 200 movies. This dataset derives from the merging of two previously independent datasets, MPII-MD (Rohrbach et al., 2015) and M-VAD (Torabi et al., 2015). The language used in audio descriptions is particularly rich of references to physical action, with respect to reference corpora (e.g. BNC corpus) (Salway, 2007).

For this reason, LSMDC16 dataset could be considered a good source of video-caption pairs of action examples, comparable to data from IMAGACT resource.

6 Experiments

Given that the objective is not to discriminate distant actions (e.g. *opening a door* vs. *taking a cup*) but rather to distinguish actions referred to by the same verb or set of verbs, the experiments herein described have been conducted on a sub-set of the LSMDC16 dataset, that have been manually annotated with the corresponding *ac*s from IMAGACT. The annotation has been carried on by one expert annotator, trained on IMAGACT conceptualization framework, and revised by a supervisor. In this way, we created a Gold Standard for the evaluation of the compared systems.

6.1 Gold Standard

The Gold Standard test set (GS) has been created by selecting one starting verb: *push*. This verb has been chosen according to the fact that, as a general action verb, it is highly frequent in the use, it applies to a high number of *ac*s in the IMAGACT Ontology (25 *ac*s) and it has a high occurrence both in IMAGACT and LSMDC16.

From the IMAGACT Ontology, all the verbs in relation of local equivalence with *push* in each of its *ac*s have been queried[8], i.e all the verbs that predicate at least one of the *ac*s linked to *push*. Then, all the captions in LSMDC16 containing one of those verbs have been manually annotated with the corresponding *ac*'s id. In total, 377 video-caption pairs have been correctly annotated[9] with 18 *ac*s, and they have been paired with 38 captions for the verbs linked to the same *ac*s in IMAGACT, consisting in a total of 14,440 similarity

[8] The verbs collected for this experiment are: *push*, *insert*, *press*, *ram*, *nudge*, *compress*, *squeeze*, *wheel*, *throw*, *shove*, *flatten*, *put*, *move*. *Move* and *put* have been excluded from this list, due to the fact that this verbs are too general and apply to a wide set of *ac*s, with the risk of introducing more noise in the computation of the similarity; *flatten* is connected to an *ac* that found no examples in LSMDC16, so it has been excluded too.

[9] Pairs with no action in the video, or pairs with a novel or difficult to assign *ac* have been excluded from the test.

judjements.

It is important to highlight that the manual annotation took into account the visual information conveyed with the captions (i.e. videos from both resources), that made possible to precisely assign the most applicable *ac* to the LSMDC16 captions.

6.2 Pre-processing of the data

As stated in the introduction, STS methods are normally tested on data within the same domain. In attempt to leverage some differences between IMAGACT and LSMDC16, basic pre-processing have been applied.

Length of caption in the two resources vary: captions in IMAGACT are *artificial*, and they only contain minimum syntactic/semantic elements to describe the *ac*; captions in LSMDC16 are transcription of more natural spoken language, and usually convey information on more than one action at the same time. For this reason, LSMDC16 captions have been splitted in shorter and simpler sentences. To do that, we parsed the original caption with StanforNLP (Qi et al., 2018), and rewrote simplified sentences by collecting all the words in a dependency relation with the targeted verbs. Table 1 shows an example of the splitting process.

FULL	*As he crashes onto the platform, someone hauls him to his feet and* **pushes** *him back towards someone.*	✓
SPLIT	*he crashes onto the platform and*	✗
	As someone hauls him to his feet	✗
	pushes *him back towards someone*	✓

Table 1: Example of the split text after processing the output of the dependency parser. From the original caption (FULL) we obtain three subcaptions (SPLIT). Only the one with the target verb is used (✓), and the rest is ignored (✗).

LSMDC16 dataset is anonymised, i.e. the pronoun *someone* is used in place of all proper names; on the contrary, captions in IMAGACT always have a proper name (e.g. John, Mary). We automatically substituted IMAGACT proper names with *someone*, to match with LSMDC16.

Finally, we also removed stop-words, which are often the first lexical elements to be pruned out from texts, prior of any computation, because they do not convey semantic information, and they sometimes introduce noise in the process. Stop-words removal has been executed in the moment of calculating the similarity between caption pairs, i.e. tokens corresponding to stop-words have been used for the representation by contextual models, but then discharged when computing sentence representation.

With these pre-processing operations, we obtained 4 variants of testing data:

- plain (LSMDC16 splitting only);
- anonIM (anonymisation of IMAGACT captions by substitution of proper names with *someone*);
- noSW (stop-words removing from both resources);
- anonIM+noSW (combination of the two previous ones).

7 Results

To benchmark the performances of the four models, we also defined a baseline that, following a binomial distribution, randomly assigns an *ac* of the GS test set (actually, baseline is calculated analytically without simulations). Parameters of the binomial are calculated from the GS test set. Table 2 shows the results at different recall@k (i.e. ratio of examples containing the correct label in the top k answers) of the three models tested.

All models show slightly better results compared to the baseline, but they are not much higher. Regarding the pre-processing, any strategy (noSW, anonIM, anonIM+noSW) seems not to make difference. We were expecting low results, given the difficulty of the task: without taking into account visual information, also for a human annotator most of those caption pairs are ambiguous.

Surprisingly, GloVe model, the only one with static pre-trained embeddings based on statistical distribution, outperforms the baseline and other contextual models by $\sim$0.2 in recall@10. It is not an exciting result, but it shows that STS with pre-trained word embedding might be effective to speed up manual annotation tasks, without any computational cost. Probably, one reason to explain the lower trend in results obtained by contextual models (BERT, USE) could be that these systems have been penalized by the splitting process of LSMDC16 captions. Example in Table

Model	Pre-processing	recall@1	recall@3	recall@5	recall@10
ONE-HOT ENCODING	plain	0.195	0.379	0.484	0.655
	noSW	0.139	0.271	0.411	0.687
	anonIM	0.197	0.4	0.482	0.624
	anonIM+noSW	0.155	0.329	0.453	0.65
GloVe	plain	0.213	0.392	0.553	**0.818**
	noSW	0.182	0.408	0.505	0.755
	anonIM	0.218	0.453	**0.568**	0.774
	anonIM+noSW	**0.279**	0.453	0.553	0.761
BERT	plain	0.245	0.439	0.539	0.632
	noSW	0.247	**0.484**	0.558	0.679
	anonIM	0.239	0.434	0.529	0.645
	anonIM+noSW	0.2	0.384	0.526	0.668
USE	plain	0.213	0.403	0.492	0.616
	noSW	0.171	0.376	0.461	0.563
	anonIM	0.239	0.471	0.561	0.666
	anonIM+noSW	0.179	0.426	0.518	0.637
Random baseline		0.120	0.309	0.447	0.658

Table 2: STS results for the models tested on IMAGACT-LSMDC scenario.

1 shows a good splitting result, while processing some other captions leads to less-natural sentence splitting, and this might influence the global result.

Model	Pre-processing	Pearson
GloVe	plain	0.336
BERT	plain	0.47
USE	plain	**0.702**

Table 3: Results on STS-benchmark.

We run similar experiments on the publicly available STS-benchmark dataset[10] (Cer et al., 2017), in order to see if the models show similar behaviour when benchmarked on a more conventional scenario. The task is similar to the one presented herein: it consists in the assessment of pairs of sentences according to their degree of semantic similarity. In this task, models are evaluated by the Pearson correlation of machine scores with human judgments. Table 3 shows the expected results: Contextual models outperform GloVe based model in a consisted way, and USE outperform the rest by large margin (about 20-30 points better overall). It confirms that model performances are task-dependent, and that results obtained in *non-conventional* scenarios can be counter-intuitive if compared to results obtained in conventional ones.

8 Conclusions and Future Work

In this paper we presented a comparison of four popular representation models (one-hot encoding, GloVe, BERT, USE) in the task of semantic textual similarity on a non-conventional scenario: action reference similarity between sentences from different domains.

In the future, we would like to extend our Gold Standard dataset, not only in terms of dimension (i.e. more LSMDC16 video-caption pairs annotated with *ac*s from IMAGACT), but also in terms of annotators. It would be interesting to observe to what extend the visual stimuli offered by video prototypes can be interpreted clearly by more than one annotator, and thus calculate the inter-annotator agreement. Moreover, we plan to extend the evaluation to other representation models as well as state-of-the-art supervised models, and see if their performances in canonical tests are confirmed on our scenario. We would also try to augment data used for this test, by exploiting dense video captioning models, i.e. videoBERT (Sun et al., 2019).

Acknowledgements

This research was partially supported by the Spanish MINECO (DeepReading RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE)), ERA-Net CHISTERA LIHLITH Project funded by Agencia Esatatal de Investigacin (AEI, Spain) projects PCIN-2017-118/AEI and PCIN-2017-085/AEI, the Basque Government (excellence research group, IT1343-19), and the NVIDIA GPU grant program.

[10] http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A pilot on semantic textual similarity. In **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, pages 385–393. Universidad del Pais Vasco, Leioa, Spain, January.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics (TACL)*, 5(1):135–146.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT - Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, 1810:arXiv:1810.04805.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Massimo Moneglia and Alessandro Panunzi. 2007. Action Predicates and the Ontology of Action across Spoken Language Corpora. In M Alcántara Plá and Th Declerk, editors, *Proceedings of the International Workshop on the Semantic Representation of Spoken Language (SRSL 2007)*, pages 51–58, Salamanca.

Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014. The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the representation of lexical encoding of Action. *LREC*, pages 3425–3432.

Massimo Moneglia. 2014. The variation of Action verbs in multilingual spontaneous speech corpora. *Spoken Corpora and Linguistic Studies*, 61:152.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(0):217 – 250.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *CoRR*, abs/1703.02507.

Alessandro Panunzi, Lorenzo Gregori, and Andrea Amelio Ravelli. 2018a. One event, many representations. mapping action concepts through visual features. In James Pustejovsky and Ielka van der Sluis, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alessandro Panunzi, Massimo Moneglia, and Lorenzo Gregori. 2018b. Action identification and local equivalence of action verbs: the annotation framework of the imagact ontology. In James Pustejovsky and Ielka van der Sluis, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Stanford University, Palo Alto, United States, January.

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

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2018. Universal Dependency Parsing from Scratch. *CoNLL Shared Task*.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for Movie Description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision*, 123(1):94–120, January.

Andrew Salway. 2007. A corpus-based analysis of audio description. In Jorge Díaz Cintas, Pilar Orero, and Aline Remael, editors, *Media for All*, pages 151–174. Leiden.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766.

Atousa Torabi, Christopher J Pal, Hugo Larochelle, and Aaron C Courville. 2015. Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. cs.CV:arXiv:1503.01070.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*.