Procesamiento del Lenguaje Natural, Revista nº 63, marzo de 2019, pp. 85-92 recibido 03-04-2019 revisado 14-05-2019 aceptado 17-05-2019

Action Type induction from multilingual lexical features

Tipos de acciones inducidas a partir de de características léxicas multilingües

Lorenzo Gregori, Rossella Varvara, Andrea Amelio Ravelli
University of Florence
{lorenzo.gregori, rossella.varvara, andreaamelio.ravelli}@unifi.it

Abstract: This paper presents a vector representation and a clustering of action concepts based on lexical features extracted from IMAGACT, a multilingual and multimodal ontology of actions in which concepts are represented through video prototypes. We computed vectors for 1,010 action concepts, where the dimensions correspond to verbs in 10 languages. Finally, an unsupervised clustering method has been applied on these data in order to discover action classes based on typological closeness. Those clusters are not language-specific or language-biased, and thus constitute an inter-linguistic classification of action domain.
Keywords: Action verbs, Vector Space Model, Multilingual Semantics, Action clustering

Resumen: Este artículo presenta una representación vectorial y un clúster de conceptos de acción basados en características léxicas extraídas de IMAGACT, una ontología de acciones multilingüe y multimodal en la que los conceptos se representan a través de prototipos de video. Calculamos vectores para 1.010 conceptos de acción, donde las dimensiones corresponden a verbos en 10 idiomas. Finalmente, se ha aplicado un método de agrupación no supervisada en estos datos para descubrir clases de acción basadas en la proximidad tipológica. Esos clústers no son específicos del idioma ni están sesgados por él, y por lo tanto constituyen una clasificación interlinguística del dominio de acción.
Palabras clave: Verbos de acción, clasificación interlinguística, agrupación de acción

1 Introduction

In the mind of a speaker, the linking between real world entities (i.e. objects and events) and their mental representation is expressed through symbols (i.e. lexical items), which are part of his own mother-tongue language[1]. This picture becomes complicated due to cognitive economy constraints (Rosch, 1978), because one lexical item productively applies to a set of objects or events, resulting in a one-to-many relations (Moneglia, 1996). Consequently, the mental representation of the world consists in a complex network of connections between entities, thoughts and lexicon. To give an idea of this complexity, we can take into account all the possible pragmatic actions that an English speaker can correctly refer to with the verb *put*. In fact, this verb activates in the mind of the speaker a series of possible events, often quite distant in terms of pragmatic execution. For example, *putting a book on the table* and *putting some jam on the bread*, from the lexical point of view of the verb, are similar events, despite their differences: in the first one, no tools are required[2], and few motoric activations are needed to perform the action; on the contrary, in the second one, a tool is needed (e.g. a knife) even if not explicitly lexicalized, and a sequence of various short actions are performed to complete the task. Moreover, only the

[1]The correlation between entities, thoughts and lexicon is often referred to as the Triangle of Reference (Ogden and Richards, 1923).

[2]According to the generativist point of view, the arm and the hand could be considered as tools (Pastra and Aloimonos, 2012).

ISSN 1135-5948. DOI 10.26342/2019-63-9

second event can be predicated correctly also by the verb *spread* which, in its turn, extends its application to a series of events, some of which are other than those activated by *put*, e.g. *people spreading around the room.*

If we extend the focus from one single speaker to a community of speakers, we can see that the mental representation is roughly shared and still holds. But what happens if we try to compare the mental representation of two or more speakers of different languages? We immediately observe a variation in the linking between items of the conceptual space and words of the lexical dimension. In fact, if we ask a Japanese speaker to lexicalize the three aforementioned actions, he would use *oku* (common translation of *put*) to predicate the event of *putting a book on the table*, *tsukeru* for *putting/spreading some jam on the bread*, and *chirabaru* for *people spreading around*, and none of these verbs are overlapping or interchangeable in their primary pragmatic predication.

These examples suggest that even a pure lexical discrimination is able to finely segment the conceptual space, and highlight closeness or distance between concepts, both in a monolingual and a multilingual setting.

Starting from these observations, the work presented in this paper is a first attempt in modelling a language-independent conceptual space representation of actions, using multilingual lexical items as a piece of evidence for action concept discrimination.

2 Connection with other works

The present work is a first attempt in solving a novel NLP task, that we will call Action Type Induction (ATI): it faces the problem of automatically identify similar action instances and group them in types. Similarly to Word Sense Induction (WSI) (Pantel and Lin, 2002), ATI can be addressed as an unsupervised clustering problem, where vectors represent actions, instead of words. This difference is not trivial, and makes ATI an intrinsically multimodal task, in which instances are video elements of performed actions, and the encoded features can be both linguistic and visual. For example, informations about verbs or textual description that refer to actions, as well as features regarding motion and trajectories, become relevant data for this task.

The combination of linguistic and visual features to perform a more accurate classification of actions has been widely used in recent years (Silberer, Ferrari, and Lapata, 2013; Hahn, Silva, and Rehg, 2019; Naha and Wang, 2016), with the development of techniques based on the integration of NLP and computer vision. Within this perspective, our work could be fruitfully exploited to build complex models for action understanding grounded on human knowledge.

Since our instances are actions, and not words, we couldn't use a co-occurrence matrix. Instead our dataset is a co-referentiality matrix, that encodes local equivalence, i.e. the ability of two verbs to refer to the same action concepts (Panunzi, Moneglia, and Gregori, 2018).

A similar approach has been used to represent typological data (Ryzhova, Kyuseva, and Paperno, 2016): a matrix of word references, in which each row corresponds to nouns from a specific semantic field and the dimensions are adjectives from different languages. An intersection of a row and a column is filled with 1 if the adjective can occur in the context and with 0 if it cannot. Given these data, they compute *typological closeness* between nouns. From their work we inherit the notion of typological closeness, that is semantic similarity based on comparison of multilingual data.

3 Action vector representation

3.1 IMAGACT

IMAGACT[3] (Moneglia et al., 2014a) is a multimodal and multilingual ontology of action that provides a video-based translation and disambiguation framework for action verbs. The resource is built on an ontology consisting in a fine-grained categorization of action concepts, 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 action concepts most commonly referred to in everyday language usage. Action concepts have been gathered through an information bootstrapping from Italian and English spontaneous spoken corpora, and the occurrences of verbs referring to physical actions have been manually annotated (Moneglia et al., 2012). Metaphorical and phraseological usages have been ex-

[3] http://www.imagact.it

cluded from the annotation process, in order to collect exclusively occurrences of physical actions.

The database evolves continuously and at present contains 10 fully-mapped languages and 17 which are underway. The insertion of new languages is obtained through competence-based extension (CBE) (Moneglia et al., 2014b) by mother-tongue speakers, using a method of ostensive definitions inspired by Wittgenstein (Wittgenstein, 1953). The informants are asked to watch each video, to list all the verbs in their language that correctly apply to the depicted action, and to provide a caption describing the event for every listed verb as an example of use.

The visual representations convey action information in a cross-linguistic environment, offering the possibility to model a conceptualization avoiding bias from monolingual-centric approaches.

3.2 Dataset

From the IMAGACT database, we derived our dataset as a binary matrix $C_{1010\times7881}$ with one row per video prototype and one column per verb for the languages considered. Matrix values are the assignments of verbs to videos made by native speakers within the CBE annotation task:

$$C_{i,j} = \begin{cases} 1 & \text{if verb } j \text{ refers to action } i \\ 0 & \text{else} \end{cases}$$

In this way, the matrix C encodes the interlinguistic lexical representation of each video prototype.

Table 1 shows the number of verbs assigned by the CBE annotators for each language. It is important to notice that the task has been performed on the whole set of 1,010 scenes for each language and the differences between the number of verbs depend on linguistic factors: some examples of verb-rich languages are (a) Polish and Serbian, in which perfective and imperfective forms are lemmatized as different dictionary entries, (b) German, that have particle verb compositionality, (c) Spanish and Portuguese, for which verbs belong to both American and European varieties.

Judgments of applicability of a verb to a video scene rely on the semantic competence of annotators. An evaluation of CBE assignments has been made for Arabic and Greek in two thesis (Mutlak, forthcoming; Mouyiaris, forthcoming); results are summarized in Table 2.

Language	**Verbs**
Arabic (Syria)	571
Danish	646
German	990
Greek	638
Hindi	470
Japanese	736
Polish	1,193
Portuguese	805
Serbian	1,096
Spanish	736
TOTAL	**7881**

Tabla 1: Number of verbs per language

Language	**Precision**	**Recall**
Arabic (Syria)	0.933	0.927
Greek	0.990	0.927

Tabla 2: Precision and Recall for CBE annotation task measured on 2 languages

3.3 Creating action vectors

In order to provide an exploitable vector representation of action prototypes, an approximated matrix C' has been created from C, by using *Singular Value Decomposition* (SVD) for dimensionality reduction.

SVD is a widely used technique in distributional semantics to reduce the feature space. The application of SVD to our dataset allowed us to obtain a fixed-size feature space, that is independent of the number of languages, and an approximation matrix, that smooths language-specific semantic differences. These results are highly desirable, considering that the number of languages in IMAGACT is growing continuously, and that the provided representation should be shared as far as possible, abstracting from lexico-semantic properties of single languages. Moreover, SVD approximation leads to some advantages in terms of computational processing, by removing the matrix sparsity and creating a relatively low dimensional space.

Finally, the output C' is a dense matrix 1010×300 that encodes lexical features of action prototypes.

4 Actions clustering

In order to obtain a language-independent classification of action concepts, we computed similarity between action prototype vectors, and then we applied an unsupervised clustering algorithm to this data. The resulting classification bypasses differences in lexicalization among languages, in favour of an *average* conceptual representation of actions. In fact, a classification based on data from only one language leads to a representation that is consistent with the semantic space segmentation operated by that language, but it may not be cross-linguistically generalized. Considering more languages together, instead, language-specific differences can be leveraged, highlighting similarities that may remain in shade if comparing monolingual classifications of actions.

Data clustering in this scenario is a complex task because we do not have any information on the number of clusters, that must be found by the clustering algorithm. Moreover, a proper evaluation of the resulting clusters is not trivial, since we need to compare one speaker's conceptual representation with the average representation resulting from summing lexical information from multiple languages.

4.1 Clustering algorithm

An ATI task can be properly considered as a variation of a Word Sense Induction (WSI) task: instead of grouping word occurrences with similar meaning based on word contexts, here we are grouping similar action occurrences based on lexical features. In this experiment we used *Affinity Propagation* (AP) (Frey and Dueck, 2007), a state-of-the-art unsupervised clustering algorithm, that has been successfully applied to accomplish WSI tasks in recent works (Alagić, Šnajder, and Padó, 2018; Arefyev, Ermolaev, and Panchenko, 2018). AP automatically identifies the optimal number of clusters for a given dataset; each cluster consists of one *exemplar* (i.e. one element of the dataset that is representative of the cluster) plus its neighbouring elements.

Results of clustering on C' matrix are summarized in Table 3.

Number of clusters	178
Min # of scenes per cluster	2
Max # of scenes per cluster	24
Average # of scenes per cluster	5.67

Tabla 3: Results of clustering algorithm

4.2 The map of action concepts

A visual map has been created for data exploration purposes, available at http://lablita.it/app/imclust/map.php. In this map each point is a cluster (i.e. a set of action videos); the spatial position of points is derived as follows:

- exemplar vectors are chosen as cluster representatives;
- the feature space has been reduced to 2 dimensions with t-SNE[4];
- this 2D representation of exemplars has been projected on x and y axes.

The action map is interactive: by clicking on each point it is possible to see the set of videos belonging to the cluster. In order to ease data exploration, cluster regions have been manually drawn and named with an English verb that roughly describes the related semantic area.

5 Evaluation

5.1 Evaluation task

To our knowledge, no similar previous evaluation work are available and, due to the peculiarity of this kind of work (i.e. language independent clustering obtained by summing of multilingual data), the possibility that a speaker of a language L may find reasonable all the clusters could be far from being positive. Nevertheless, in order to evaluate the obtained Action Type clusters and the applied methodology, we designed a task to perform an evaluation based on human percepts on scenes similarity.

The evaluation consists in a two-alternative forced-choice similarity task in which, given as target a scene s from a cluster c, participants were asked to chose the most similar scene to the target among two other scenes. One of the scene used for comparison belongs to the same cluster c of the target, whereas the other was selected among scenes not belonging to it. We expected annotators to choose the scene belonging

[4]t-SNE is a dimensionality reduction algorithm, specifically designed for visual representation of high dimensional data (Maaten and Hinton, 2008).

to the same cluster as more similar to the target one. If human judgments mirror our unsupervised clusters, we interpret it as a reliability result. Figure 1 shows one item of the evaluation test: annotators judge which video, between 1 and 2, is more similar to the target.

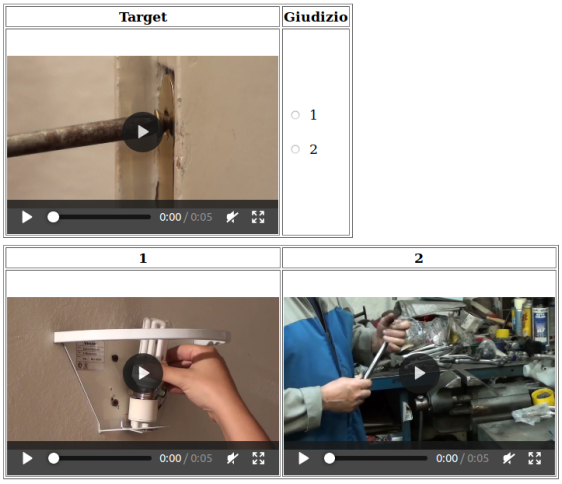

Figura 1: One item of the evaluation test

We conducted a preliminary evaluation on 11 randomly selected clusters. These were selected considering the number of scenes they contain: the clusters considered were in the middle range for number of scenes (one standard deviation around the mean[5]).

We presented to participants every possible couple of scenes belonging to the same cluster, thus resulting in a dataset of 125 datapoints. The third scene of each evaluation item was randomly selected among scenes not belonging to the considered cluster. Moreover, we checked for the similarity distance of this scene from the cluster, and prepared two version of the test: the third scene was alternatively selected among the 15th most similar scenes not belonging to the target cluster or among other farther scenes, and this selection was inverted in the second version of the test. In this way, we can evaluate finely the precision in the categorization of closer concepts. Furthermore, we divided each test in two, in order to avoid a very time consuming test. As a result we obtained 2 different tests with 2 alternative version each (4 tests in total).

[5]We preferred clusters with a number of scenes in the middle range for two reasons: 1- to avoid outlier clusters; 2- to keep the number of items per test small enough, feasible in less then 30 minutes

The first test was performed by 10 annotators, 5 for each version, while the second test was performed by 6 annotators, 3 for each version. All the annotators are Italian native speakers. Note that Italian is not among the languages included in our multilingual matrix. We believe this is an additional strength of our evaluation, because results are not biased from the presence of annotators' mother language into the multilingual matrix.

This preliminary evaluation is meant as a first check on the task suitability. An extensive evaluation with a crowd sourcing platform has been planned, which will consider a larger number of clusters and will collect judgments from speakers of different languages. It will be interesting to analyze the influence of participants mother language in the results, observing language-specific differences.

5.2 Evaluation results

Tables 4 and 5 show results of Test 1 for each annotator; tables 6 and 7 show results of Test 2 for each annotator. Values range from 0 to 1, where 1 indicates that the scenes judged as more similar to the target correspond to the cluster's internal ones. The two versions of the test had a small difference in the percentage of correct pairs (0.88 for Test 1.1, 0.91 for test 1.2 and 0.97 for Test 2.1, 0.92 for Test 2.2), suggesting us that the choice of the third (not belonging to the cluster) scene is relevant. Moreover, results show a small difference between near and far scenes: if the third scene was farther to the cluster, annotators judged the intra-cluster scene as similar to the target scene more easily. Finally Cohen's k has been measured between each annotator and the automatic clustering assignments; the resulting agreement is high: 0.75 for Test 1.1; 0.81 for Test 1.2; 0.83 for Test 2.1; 0.80 for Test 2.2.

	near	**far**	**total**	***k***
annot 1	0.80	0.92	0.86	0.84
annot 2	0.80	0.88	0.84	0.71
annot 3	0.80	0.92	0.86	0.80
annot 4	0.96	0.88	0.92	0.67
annot 5	0.88	0.92	0.90	0.72
Average	**0.85**	**0.90**	**0.88**	**0.75**

Tabla 4: Evaluation results - Test 1.1

Table 8 and 9 report the distribution of shared judgments among participants, i.e. how many items are evaluated according to

	near	far	total	*k*
annot 6	0.88	0.92	0.90	0.84
annot 7	0.92	0.92	0.92	0.84
annot 8	0.88	0.92	0.90	0.80
annot 9	0.88	0.92	0.90	0.79
annot 10	0.92	0.92	0.92	0.79
Average	**0.89**	**0.92**	**0.91**	**0.81**

Tabla 5: Evaluation results - Test 1.2

	near	far	total	*k*
annot 11	0.97	0.97	0.97	0.88
annot 12	0.97	1.00	0.98	0.74
annot 13	0.97	0.97	0.97	0.86
Average	**0.97**	**0.98**	**0.97**	**0.83**

Tabla 6: Evaluation results - Test 2.1

	near	far	total	*k*
annot 14	0.88	1.00	0.93	0.76
annot 15	0.91	0.95	0.93	0.87
annot 16	0.88	0.97	0.92	0.78
Average	**0.89**	**0.97**	**0.92**	**0.80**

Tabla 7: Evaluation results - Test 2.2

the clustering and by how many annotators. For example, the first row of Table 8 reports the number of evaluated items where all of the 5 annotators (5/5) identified the scene belonging to the cluster as more similar to the target.

In general, the evaluation confirmed the validity of the cluster obtained: clusters seem to correspond to what we call Action Type, since they mirror with a high percentage to human judgments. The matrix and the clusters obtained can be, thus, interpreted as interlinguistic and cognitively valid representation of Action Types. With further evaluations we want to investigate to which extent linguistic (lexical) representation mirrors cognitive categorization, thus providing new insights on how we organize concepts.

6 Conclusion and future works

This work describes an experiment of Action Type Induction, that has been addressed through an unsupervised clustering on a matrix of multilingual lexical features. Vectors encode the information about the applicability of verbs (in 10 languages) to action concepts and are extracted from the IMAGACT Ontology of Action. Affinity Propagation clustering algorithm has been applied to these data and the 1,010 video scenes have been grouped in 178 clusters, that represent a conceptualization of action domain. A preliminary evaluation has been made on a set of 11 clusters by 16 annotators, obtaining encouraging results. The very next step will be the implementation of an evaluation campaign to obtain a reliable measure of clustering accuracy without a gold standard.

Pair agr.	Test 1.1	Test 1.2	Total
5/5	39	40	79
4/5	2	4	6
3/5	3	2	5
2/5	2	1	3
1/5	3	3	6
0/5	1	0	1

Tabla 8: Agreement summary on scenes internal to the target cluster in Test 1

Pair agr.	Test 2.1	Test 2.2	Total
3/3	71	67	138
2/3	1	3	4
1/3	2	1	3
0/3	0	3	3

Tabla 9: Agreement summary on scenes internal to the target cluster in Test 2

The action representation provided in this experiment is based on lexical features only, but ATI tasks can exploit several kinds of action features. For this reason, this work can be considered as a first step in the creation of a more complete vector representation of actions that encodes other aspects of action domain. Some examples of features that, in principle, could enrich these vectors regard event structures (e.g. goal, semantic frames), linguistic properties of linked verbs (e.g. thematic roles, aktionsart), visual properties of the scenes (e.g. action trajectory, spatial relation among objects), motor description of performed movements.

References

Alagić, D., J. Šnajder, and S. Padó. 2018. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Arefyev, N., P. Ermolaev, and A. Panchenko. 2018. How much does a word weigh? weighting word embeddings for word sense induction. *arXiv preprint arXiv:1805.09209*.

Frey, B. J. and D. Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Hahn, M., A. Silva, and J. M. Rehg. 2019. Action2Vec: A Crossmodal Embedding Approach to Action Learning. *arXiv.org*, page arXiv:1901.00484, January.

Maaten, L. v. d. and G. Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Moneglia, M. 1996. Prototypical vs. non-prototypical predicates: ways of understanding and the semantic partition of lexical meaning. *QDLF - Quaderni del Dipartimento di Linguistica*, 7:163–181.

Moneglia, M., S. Brown, F. Frontini, G. Gagliardi, F. Khan, M. Monachini, and A. Panunzi. 2014a. The imagact visual ontology. an extendable multilingual infrastructure for the representation of lexical encoding of action. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Moneglia, M., S. Brown, A. Kar, A. Kumar, A. K. Ojha, H. Mello, Niharika, G. N. Jha, B. Ray, and A. Sharma. 2014b. Mapping Indian Languages onto the IMAGACT Visual Ontology of Action. In G. N. Jha, K. Bali, S. L, and E. Banerjee, editors, *Proceedings of WILDRE2 - 2nd Workshop on Indian Language Data: Resources and Evaluation at LREC'14*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Moneglia, M., F. Frontini, G. Gagliardi, I. Russo, A. Panunzi, and M. Monachini. 2012. Imagact: deriving an action ontology from spoken corpora. *Proceedings of the Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-8)*, pages 42–47.

Mouyiaris, A. forthcoming. I verbi d'azione del greco nell'ontologia IMAGACT. Master's thesis.

Mutlak, M. forthcoming. *I verbi di azione dell'arabo standard nell'ontologia dell'azione IMAGACT*. Ph.D. thesis.

Naha, S. and Y. Wang. 2016. Beyond verbs: Understanding actions in videos with text. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 1833–1838. IEEE.

Ogden, C. and I. A. Richards. 1923. The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism. *8th ed. 1923. Reprint New York: Harcourt Brace Jovanovich.*

Pantel, P. and D. Lin. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619. ACM.

Panunzi, A., M. Moneglia, and L. Gregori. 2018. Action identification and local equivalence of action verbs: the annotation framework of the imagact ontology. In J. Pustejovsky and I. van der Sluis, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Pastra, K. and Y. Aloimonos. 2012. The minimalist grammar of action. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1585):103–117.

Rosch, E. 1978. Principles of Categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*. Lawrence Elbaum Associates.

Ryzhova, D., M. Kyuseva, and D. Paperno. 2016. Typology of adjectives benchmark for compositional distributional models. In *Proceedings of the 10th Language Resources and Evaluation Conference*, pages 1253–1257.

Silberer, C., V. Ferrari, and M. Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582. Association for Computational Linguistics.

Wittgenstein, L. 1953. *Philosophische Untersuchungen*. Suhrkamp Verlag.