Linking IMAGACT ontology to BabelNet through action videos

Lorenzo Gregori
University of Florence
lorenzo.gregori@unifi.it

Alessandro Panunzi
University of Florence
alessandro.panunzi@unifi.it

Andrea Amelio Ravelli
University of Florence
arameliorgmail.com

Abstract

English. Herein we present a study dealing with the linking of two multilingual and multimedia resources, BabelNet and IMAGACT, which seeks to connect videos contained in the IMAGACT Ontology of Actions with related action concepts in BabelNet. The linking is based on a machine learning algorithm that exploits the lexical information of the two resources. The algorithm has been firstly trained and tested on a manually annotated dataset and then it was run on all the data, allowing to connect 773 IMAGACT action videos with 517 BabelNet synsets. This linkage aims to enrich BabelNet verbal entries with a visual representations and to connect the IMAGACT ontology to the huge BabelNet semantic network.

Italiano. *In questo articolo si presenta uno studio sul linking tra due risorse linguistiche multilingui e multimediali, BabelNet e IMAGACT. L'esperimento ha l'obiettivo di collegare i video dell'ontologia dell'azione IMAGACT con i concetti azionali contenuti in BabelNet. Il collegamento è realizzato attraverso un algoritmo di Machine Learning che sfrutta l'informazione lessicale delle due risorse. L'algoritmo è stato addestrato e valutato su un dataset annotato manualmente e poi eseguito sull'insieme totale dei dati, permettendo di collegare 773 video di IMAGACT con 517 synset di BabelNet. Questo linking ha lo scopo di arricchire le entrate verbali di BabelNet con una rappresentazione visuale e di collegare IMAGACT alla rete semantica di BabelNet.*

1 Introduction[1]

Ontologies are widely used to represent language resources on the web, allowing them to be easily accessed and exploited by machines. For this reason, data interconnection between different semantic resources is a crucial task in order to enhance disambiguation and information retrieval capabilities in Artificial Intelligence, as evidenced by the increasing research into mapping and linking techniques among ontologies (Otero-Cerdeira et al., 2015). Nevertheless, ontology mapping has to face the problem of concept representation mismatch between resources, due to different building criteria and purposes (Siemoneit et al., 2015). Instance matching techniques play an important role in this context, allowing to connect entities from heterogeneous data resources which refer to the same real-world object (Castano et al., 2008; Nath et al., 2014).

Aside the general interest for knowledge bases interconnection in a web-based perspective, there is also a growing interest in multimodal resources, which combine textual and visual data. These resources can be exploited by intelligent algorithms integrating vision and natural language processing techinques[2]. This integrated approach was successfully applied for some challenging tasks involving verbs and their action reference as a video. Regneri et al. (2013) developed machine learning models for the automatic identification of similarity among actions, by using a corpus of natural language descriptions, derived from the videos of the *MPII Cooking Composite Activities* dataset, which represents actions involved in basic cooking tasks. Instead, the algorithm developed by Mathe

[1] Lorenzo Gregori developed the linking algorithm and wrote sections 3, 4, and 5; Andrea Amelio Ravelli performed the data annotation and wrote sections 1 and 2; Alessandro Panunzi supervised the research work and revised the paper.

[2] Several works in this field have been developed within The European Network on Integrating Vision and Language (iV&L Net), http://ivl-net.eu/

et al. (2008) extracts higher level semantic features in common among a sample set of verbs, using a fine-grained analysis of the represented action concepts, intended as a subsequent stable set of abstract features of the objects involved in the videos. Within this interdisciplinary perspective, a knowledge base which relates verbal lemmas in different languages with video prototypes can help in serveral applications, and be exploited by both humans and machines.

2 Resources

This paper presents a linking between BabelNet (Navigli and Ponzetto, 2012a) and IMAGACT (Moneglia et al., 2014a), two multilanguage and multimedia resources suitable for automatic translation and disambiguation tasks (Russo et al., 2013; Moneglia, 2014; Moro and Navigli, 2015).

2.1 BabelNet

BabelNet[3] is a multilingual semantic network created from the mapping together of the WordNet thesaurus and the Wikipedia enciclopedia. At present, BabelNet 3.7 contains 271 languages and it is the widest multilingual resources available for semantic disambiguation. Concepts and entities are represented by BabelSynsets (BS), extensions of WordNet synsets: a BS is a unitary concept identified by several kinds of informations (semantic features, glosses, usage examples, etc.) and related to lemmas (in any language) which have a sense matching with that concept. BSs are not isolated, but connected together by semantic relations. Moreover, BabelNet received a large contributions from its mapping with other resources such as ImageNet, GeoNames, OmegaWiki (along with many others), which increased its information beyond the lexicon and produced a wide-ranging, multimedia knowledge base.

2.2 IMAGACT

IMAGACT[4] is a visual ontology of action that provides a video-based translation and disambiguation framework for general verbs. The database evolves continously (Moneglia et al., 2014b) and at present contains 9 fully-mapped languages and 13 which are underway. The resource is built on an ontology containing a fine-grained categorization of action concepts, each represented by one or more video prototypes as recorded scenes and 3D animations. IMAGACT currently contains 1,010 scenes which encompass the action concepts most commonly referred to in everyday language usage[5]. The links between verbs and video scenes are based on the co-referentiality of different verbs with respect to the action expressed by a scene (i.e. different verbs can describe the same action, visualised in the scene). The visual representations convey the action information in a cross-linguistic environment and IMAGACT may thus be exploited for reference disambiguation in automatic and assisted translation tasks (Panunzi et al., 2014).

3 Related works

Other attempts have previously been made to link IMAGACT with other resources. Two experiments by De Felice et al. (2014) and by Bartolini et al. (2014) were conducted in an intralinguistic perspective: their aim was to evaluate the results of a mapping between the action concepts defined in ItalWordNet and the ones categorized by IMAGACT (in terms of perfect matches or hypernym/hyponym relations).

On the contrary, the objective behind our work is to obtain a light link between the resources by enriching the action concepts in BabelNet with a visual representation; in this way, we overpass the problem of finding a match between the generic semantic concepts in BabelNet and the specific pragmatic concepts in IMAGACT. This methodology is also enforced by the multilingual frame in which the experiment is conducted. As a matter of fact, the relation between words and concepts can deeply differ across languages, while the prototypical scenes ensure a language-independent modality which is able to keep together the different lexicalizations of the action space.

This work is a further step from a previous IMAGACT-BabelNet linking experiment (Gregori et al., 2015). Even if it was just a feasibility test to check the consistency of the linking, we reported good results in automatic assignment of IMAGACT prototypical scenes to BabelNet synsets. For this reason, we built a bigger dataset and we went from a metric-based to a Machine Learning algorithm to be run on the whole set of IMAGACT

[3] http://babelnet.org

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

[5] The data is derived from the annotation of verb occurrences in spontaneous spoken corpora (Moneglia et al., 2012)

scenes.

4 Linking experiment

The aim of this experiment is to link the IMAGACT video scenes to the BabelNet interlinguistic concepts (BabelSynsets). In fact, the BabelNet objects are already enriched with visual objects, though this information contains static images which are inadequate for representing action concepts. In this way, adding video scenes to the verbs is very desirable and would suggest itself as a natural extension of BabelNet.

4.1 Training and test set

A manually annotated dataset of 50 scenes and 57 BabelSynsets (2,850 judgments) was created in order to test the algorithm and evaluate the results.

The sampling was carried on in two steps. First of all, a purely actional semantic area has been selected by taking BSs and scenes linked to 7 English action verbs, which are general and very frequent in the language use: *put*, *move*, *take*, *insert*, *press*, *give* and *strike*. The wide variation of these verbs allowed us to obtain a big set of concepts, with a high variation in terms of frequency and generality. On this set, a second sampling has been performed by preserving the variability in terms of number of connected verbs, that is a measurable parameter in both the resources.

Each $\langle$BS,Scene$\rangle$ pair has been evaluated to check if the scene is appropriate in representing the BS. Three annotators compiled the binary judgment table and we reported the values shared by at least 2 of 3. The measured Fleiss' kappa inter-rater agreement for this task was 0.74 [6].

Finally, the dataset has been split in a training set and a test set, with the proportions of 80% and 20% respectively (10 randomly chosen scenes for the test set and the remaining 40 scenes for the training set).

4.2 Algorithm

For this task, we developed a new algorithm which uses Machine Learning techniques, by exploiting the training set. As in the previous experiment, the features are extracted from the lexical items belonging to both the candidate BabelSynset and its neighbours[7]. Beside the algorithm, a baseline is determined by calculating the ratio $\frac{nsb}{nb+ns}$ for each pair and setting a threshold of 0.04, that maximizes the F-measure on our dataset.

Table 1 reports on the 17 languages common to both BabelNet and IMAGACT, detailing the relative number of verbs in each, and constitutes the quantitative data which the matching algorithms can exploit.

Language	BN Verbs	IM Verbs
English (EN)	29,738	1,299
Polish (PL)	9,660	1,193
Chinese (ZH)	9,507	1,171
Italian (IT)	7,184	1,100
Spanish (ES)	6,159	736
Russian (RU)	4,975	34
Portuguese (PT)	4,624	776
Arabic (AR)	3,738	804
German (DE)	3,754	992
Norwegian (NO)	1,729	115
Danish (DA)	1,685	646
Hebrew (HE)	1,647	160
Serbian (SR)	858	1,124
Hindi (HI)	831	466
Urdu (UR)	233	78
Sanskrit (SA)	33	276
Oriya (OR)	6	160
Total	86,361	11,130

Table 1: The 17 shared languages of BabelNet (BN) and IMAGACT (IM) with verbal lemma counts.

The basic features that we used for this experiment are:

- ns: the number of verbs connected to the Scene;
- nb: the number of verbs connected to the BS;
- nsb: the number of verbs that are shared between the Scene and the BS;

These 3 features have been calculated for each candidate BS and for the ones which are semantically related to it. We took the 8 BabelNet semantic relations available for verbs (see table 2) and for each BS we extracted 8 groups of related synsets, each one containing the set of BS connected to the main one by the same relation. Then, ns, nb and nsb are calculated for each group by summing the values of the BSs belonging to it.

[6] The manually annotated training set is published at http://bit.ly/29J0ypx

[7] This test is based on BabelNet 3.6; the data was extracted using the Java API (Navigli and Ponzetto, 2012b)

The feature set is comprised of 27 features: 3 features for the main BS and 3 features for each BabelNet relation. The set of candidates consists of all the possible BSs for each verb connected to the scene. A machine learning algorithm was trained on the annotated dataset: we used Support Vector Machine (SVM) classifier with a RBF kernel.

Table 2 shows the list of relations between the verbal BSs ranked by their relevance values for this task; this value is measured with Information Gain on the annotated dataset.

BabelNet relations	***IG* value**
Hyponym	0.057
Hypernym	0.026
Also See	0.019
Verb Group	0.019
Gloss Related	0.009
Entailment	0.003
Antonym	0.000
Cause	0.000

Table 2: Relations between verbal BSs.

4.3 Results

The algorithm was run on the training set and evaluated on the test set; the results are reported in Table 3.

	Baseline	**ML Algorithm**
	$th = 0.04$	27 features
Pr	0.580	0.833
Re	0.529	0.441
Fm	0.553	0.577

Table 3: Precision, Recall and F-measure of BSs to scenes linking task calculated on the test set for the algorithm and the baseline.

The results in terms of F-measure are not so satisfying and the value obtained with the algorithm is barely better than the baseline. Despite this, it's important to consider the difference with the baseline in terms of precision and recall, since precision is more important for this task: for this reason, the algorithm provides a much more reliable result compared to the baseline. We also have to point out that a low recall is mainly caused by multiple possiblities in the interpretation of a scene from different points of view: for example, the scene linked to the English verb *to throw* described by the sentence *John throws the ball to Mark* can represent not only a sense of *throw*, but also senses of other verbs, like *to play* or *to catch*, that refer to different semantic concepts; in these cases, the scene in IMAGACT is not linked to the alternative verbs, but it can be described with them (i.e. *John and Mark play with the ball*, *Mark catches the ball*). For this reason, the manual annotation provides more BS-to-scene relations than an algorithm can foresee on the basis of a lexical match, causing a low recall value.

Table 4 reports some statistics about the linking process; the entire results are browsable at the page `http://bit.ly/2a4FefT`.

IM Scenes linked to BS	773
BS linked to Scenes	517
IM English Verbs related to Scenes	544
BabelNet English Verbs related to BS	1,100

Table 4: IMAGACT-BabelNet linking numbers

Switching to Machine Learning had a strong impact on this linking task. The main advantage from the previous linking experiment (Gregori et al., 2015) is that now the number of BSs that can be assigned to each scene is variable, depending on the different reference possibilities that the BSs have. This is coherent with the BabelNet structure where we find very general concepts, that can be represented by several action prototypes, and specific ones, for which one prototype is enough to provide a clear representation.

For example the BS ”bn:00090224v” (*Put into a certain place or abstract location*) expresses a general concept and is linked to 72 scenes, comprising the actions involving one or more objects or a body part, relating to different ways of *putting* (like *inserting*, *throwing*, *attaching*,...) or to different states of the Theme (e.g. solid or liquid). Conversely, the BS ”bn:00084326v” (*Fasten with buttons*) is much more specific and is linked to only one scene (*c17d7346*) that represents a man that fastens his jacket.

5 Conclusions

The experiment described in this paper shows that is possible to obtain an extensive linking between IMAGACT and BabelNet through visual entities (see Figure 1 for a visual representation of a linking example); this can be advantageous for both the resources. BabelNet can add a clear video representation of the verbal synsets that refer to

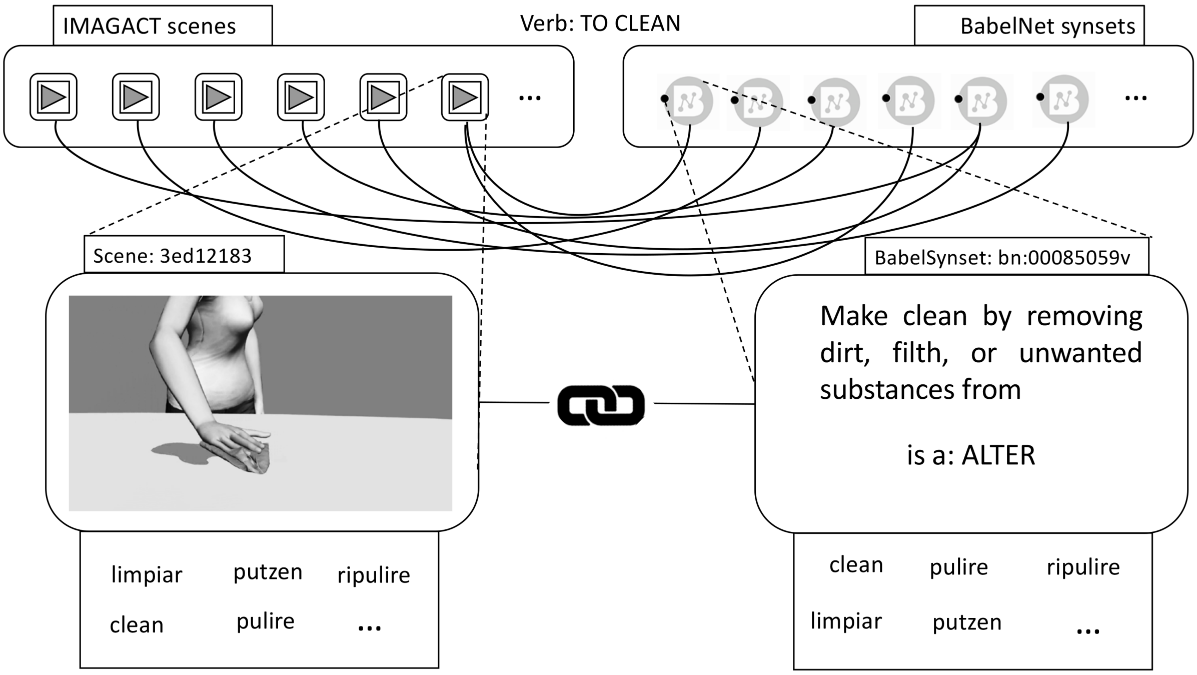

Figure 1: IMAGACT scene to BabelNet synset linking example

actions; IMAGACT can import verb translation candidates from many languages by exploiting the BabelNet semantic network; their integration can be exploited as a unified multimedial resource to accomplish complex tasks that combine natural language processing and computer vision.

Finally, we feel important to note that this procedure is scalable and the statistical model can be retrained at resource changes. This is a fundamental feature, especially considering the continuous update of IMAGACT languages and lemmas inventory.

Acknowledgments

This research has been supported by the MODELACT Project, funded by the Futuro in Ricerca 2012 programme (Project Code RBFR12C608); http://modelact.lablita.it.

References

[Bartolini et al.2014] Roberto Bartolini, Valeria Quochi, Irene De Felice, Irene Russo, and Monica Monachini. 2014. From synsets to videos: Enriching italwordnet multimodally. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

[Castano et al.2008] S. Castano, A. Ferrara, D. Lorusso, and S. Montanelli. 2008. On the ontology instance matching problem. In *Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on*, pages 180–184, Sept.

[De Felice et al.2014] Irene De Felice, Roberto Bartolini, Irene Russo, Valeria Quochi, and Monica Monachini. 2014. Evaluating ImagAct-WordNet mapping for English and Italian through videos. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume I, pages 128–131. Pisa University Press.

[Gregori et al.2015] Lorenzo Gregori, Andrea Amelio Ravelli, and Alessandro Panunzi. 2015. Linking dei contenuti multimediali tra ontologie multilingui: i verbi di azione tra imagact e babelnet. In C. Bosco, F.M. Zanzotto, and S. Tonelli, editors, *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 150–154. Accademia University Press.

[Mathe et al.2008] S. Mathe, A. Fazly, S. Dickinson, and S. Stevenson. 2008. Learning the abstract motion semantics of verbs from captioned videos. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, June.

[Moneglia et al.2012] Massimo Moneglia, Francesca Frontini, Gloria Gagliardi, Irene Russo, Alessandro

Panunzi, and Monica 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.

[Moneglia et al.2014a] Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014a. The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

[Moneglia et al.2014b] Massimo Moneglia, Susan Brown, Aniruddha Kar, Anand Kumar, Atul Kumar Ojha, Heliana Mello, Niharika, Girish Nath Jha, Bhaskar Ray, and Annu Sharma. 2014b. Mapping Indian Languages onto the IMAGACT Visual Ontology of Action. In Girish Nath Jha, Kalika Bali, Sobha L, and Esha 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).

[Moneglia2014] Massimo Moneglia. 2014. Natural Language Ontology of Action: A Gap with Huge Consequences for Natural Language Understanding and Machine Translation. In Zygmunt Vetulani and Joseph Mariani, editors, *Human Language Technology Challenges for Computer Science and Linguistics*, volume 8387 of *Lecture Notes in Computer Science*, pages 379–395. Springer International Publishing.

[Moro and Navigli2015] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June. Association for Computational Linguistics.

[Nath et al.2014] Rudra Nath, Hanif Seddiqui, and Masaki Aono. 2014. An efficient and scalable approach for ontology instance matching. *Journal of Computers*, 9(8).

[Navigli and Ponzetto2012a] Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Navigli and Ponzetto2012b] Roberto Navigli and Simone Paolo Ponzetto. 2012b. Multilingual WSD with just a few lines of code: the BabelNet API. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Korea.

[Otero-Cerdeira et al.2015] Lorena Otero-Cerdeira, Francisco J. Rodrguez-Martnez, and Alma Gmez-Rodrguez. 2015. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949 – 971.

[Panunzi et al.2014] Alessandro Panunzi, Irene De Felice, Lorenzo Gregori, Stefano Jacoviello, Monica Monachini, Massimo Moneglia, Valeria Quochi, and Irene Russo. 2014. Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology. In *XVI EURALEX International Congress: The User in Focus*, pages 1163–1170, Bolzano / Bozen, 7/2014. EURALEX 2014, EURALEX 2014.

[Regneri et al.2013] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

[Russo et al.2013] Irene Russo, Francesca Frontini, Irene De Felice, Fahad Khan, and Monica Monachini. 2013. Disambiguation of Basic Action Types through Nouns Telic Qualia. In Roser Saur, Nicoletta Calzolari, Chu-Ren Huang, Alessandro Lenci, Monica Monachini, and James Pustejovsky, editors, *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon. Generative Lexicon and Distributional Semantics*, pages 70–75.

[Siemoneit et al.2015] Benjamin Siemoneit, John Philip McCrae, and Philipp Cimiano. 2015. Linking four heterogeneous language resources as linked data. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 59–63, Beijing, China, July. Association for Computational Linguistics.