Learning Grasping Possibilities for Artifacts

Irene Russo
Istituto di linguistica Computazionale “A . Zampolli” CNR
irene.russo@ilc.cnr.it

Abstract

English. In this paper we want to test how grasping possibilities for concrete objects can be automatically classified. To discriminate between objects that can be manipulated with one hand and the ones that require two hands, we combine conceptual knowledge about the situational properties of the objects, which can be modeled with distributional semantic approaches, and physical properties of the objects (i.e. their dimensions and their weights), which can be found in the web through crawling.

Italiano. *In questo articolo vogliamo testare come le possibilit di manipolazione degli oggetti concreti possano essere classificate automaticamente. Per distinguere tra oggetti che possono essere manipolati con una mano e oggetti che richiedono due mani, combiniamo conoscenza concettuale sulle proprietà situazionali dell’oggetto - rappresentandola secondo il paradigma della semantica disribuzionale - con le proprietà fisiche degli oggetti (le loro dimensioni e il loro peso) estratte dal web mediante crawling.*

1 Introduction

Distributional semantic models of word meanings are based on representations that want to be cognitively plausible and that, as a matter of fact, have been tested to produce results correlated with human judgments when concepts similarity and automatic conceptual categorizations are the aim of the experiment (Erk, 2012; Turney and Pantel, 2010).

These approaches share the idea that two nominal concepts are similar and can be clustered in the same group if the corresponding lexemes occur in comparable linguistic contexts.

Their success is also due to the expectations of the Natural Language Processing (henceforth NLP) community: both for count and predictive models of distributional semantics (Baroni et al. 2014), the core idea is that encyclopedic knowledge packed in a big corpus can improve the performance in tasks such as word sense disambiguation.

However, purely textual representations turn out to be incomplete because, in language learning and processing, human beings are exposed to perceptual stimuli paired with linguistic ones; the old AI dream to ground language in the world requires the mapping between these two sources of knowledge. Can distributional representations of concrete nouns be helpful for the automatic classification of objects, when grasping possibilities are the focus? Could they help to discriminate between objects that can be manipulated with one hand and the ones that require two hands? More generally, how much knowledge about the physical world can be found in language?

Inspired by the cognitive psychology literature on the topic, in this paper artifactual categories are theorized as situated conceptualization where physical and situational properties meet (Barsalou 2002). Situational properties describe a physical setting or event in which the target object occurs (as *grocery store, fruit basket, slicing, picnic* for *apple*). In an action-based categorization of objects, these kinds of properties function as a complex relational system, which links the physical structure of the object, its use, the background settings, and the design history (Chaigneau et al. 2004). Situational properties can be derived from distributional semantic models, where each co-occurrence vector approximates the encyclopedic knowledge about its referent.

A complementary, but more action-oriented idea, is the psychological notion of affordance as the possibilities for actions that every environmental object offers (Gibson 1979). Conceptual information concerning objects affordances can be partially acquired through language, considering verb-direct object pairs as the linguistic realizations of the relations between the actions that can be performed by an agent, and the objects involved in those actions. Affordance verbs, intended as verbs that select a distinctive action for a specific object, can be discovered through statistical measures in corpora (Russo et al. 2013).

In this paper we want to test which source of knowledge is better for classifying artifacts grasping possibilities.

The main assumption of this paper is that the primary affordance for grasping of an artifact largely depends on its physical properties, in particular dimensions and weight. Such features are generally specified in e-commerce websites. Extracting these values for many similar items, for example for all instances of "plate", may help to automatically represent average dimensions and variability for that object.

The paper is structured as follow: section 2 reports on the manual annotation of grasping possibilities for a set of 143 artifacts, discussing the definition of the gold standard that will be the dataset for classification experiments in section 3. Section 4 presents conclusions and ideas for future work.

2 Manual Annotation of Grasping Possibilities

Concerning grasping possibilities for concrete objects, we expect as relevant several features. First of all, objects dimensions strongly influence the type of grasp afforded by objects. For instance, we are likely to grasp a tennis ball with a whole hand, but a soccer ball with two hands: the difference between the two spheres clearly is in their diameter.

However, if the soccer ball is made of foam, we might be induced to use a single hand instead of two; therefore, objects constituency is another parameter that influences the type of grasp afforded. Moreover, heavy objects require a type of grasp different from the one required by the light ones. Apart from these features, we should also consider more subjective factors, such as culture, past experience with objects, or intentions. This is particularly evident for artifacts and tools, that are the kind of objects most typically involved in manipulation and grasping and that often have a part that is specifically designed (or more suited than others) for grasping, for its shape and conformation, such as a handle (which we may call affording parts; cf. De Felice, 2015; in press). However, such parts (e.g. the handle of a cup) are usually grasped when the agents intention is to use the object for its canonical function (e.g. to drink from the cup), whereas in other cases it may be ignored and a different grasp could be performed (e.g. the whole cup might be taken from the above if we simply wanted to displace it).

Therefore, we can individuate, at least, four different grasp types afforded by concrete entities (cf. infra): the undifferentiated one-handed or two-handed grasps; a grasp by part, i.e. directed to a specific part of the object; a grasp with instrument, for substances, aggregates or every sort of things usually manipulated with some other object.

In order to obtain a gold standard annotation of artifacts grasp possibilities, we first searched WordNet 3.0 for all the nouns that have artifact as hyperonym, obtaining a list of 1510 synsets. From this list, we chose the nouns that have enough pictures as products sold on amazon.com, since it was our intention to extract objects dimensions from this website for classification experiments (cf. 3). We crawled amazon.com extracting the first 15 pages resulting from a search based on keywords (i.e. "mug", "bottle", etc.). Since some noise it's possible after looking at products pictures, we selected the nouns for which at least 15 pages about that object sold on amazon.com were homogeneous - i.e. they contain objects of the same type-. We obtained a total number of 143 nouns. Then, for each of these nouns, we manually annotated the type of grasp (also more than one) afforded by the object, according to the following classes:

- One-handed grasp: this kind of grasp is for objects that have no handles or protruding parts suited for the grasp, and that can be grasped by using only one hand. The size of two of the objects dimensions (length, width or thickness) usually does not exceed the maximum span of a hand with at least two fingers bent in order to grasp and hold something. E.g.: bowl, bottle, candle, shell, necklace, clothes peg.
- Two-handed grasp: this kind of grasp is for objects that have no handles or protruding

Table 1: Number of items per classes in the gold standard.

class	#nouns
onehand	43
onehandpart	1
onetwohand	25
part	23
twohand	73
twohandpart	3

parts suited for the grasp, and that are usually grasped with two hands, because their size exceeds the maximum span of a single hand. E.g.: board, soccer ball, player piano, table, computer.

- Grasp by part: this kind of grasp is for: (i) small or large objects that have a part specifically designed for the grasping; (ii) entities that have a well identifiable part that, even if it is not specifically designed for this specific purpose, is more suited than others for the grasping thanks to its shape and conformation. E.g. knife, jug, axe, trolley, bag.

- Grasp with instrument: this kind of grasp is mainly for substances, aggregates, and entities which cannot be (or are usually not) controlled without using some other object (an instrument, generally a container). E.g. water, broth, flour, bran, sand.

The dataset of 143 nouns have been annotated by two annotators and the inter-annotator agreement was 0.66. Since we need a gold standard for experiments, we managed disagreements reaching a consensus on every noun.

The gold standard contains items assigned to 6 classes, distributed as in Table 1.

3 Semantic and physical knowledge about artifacts: guessing grasping possibilities

The way humans can grasp an object can be designed as a function that depends on multiple variables, such as the presence of affording parts (i.e. handle for bag), its shape, its dimensions, its weight and the final aim of the action of grasping. In this paper we want to test which one of these features can help in classifying artifacts that have been manually annotated according to 6 categories (see par. 2). In particular we experiment with a combination of 4 features provided for each noun:

- distributional semantics information from two corpora (GoogleNews and instructables.com) obtained with word2vec toolkit (Mikolov et al. 2013;

- average dimensions (height, length and depth) for each object crawling at least 15 pages per object from amazon.com;

- average weight for each object crawling at least 15 pages per object from amazon.com;

- co-occurrence matrix in the corpus instructables.com with nouns that are affording parts, extracting the syntactic pattern AFFORDING PART NOUN of ARTIFACT (e.g. ”handle of the bag”).

Because all the big corpora available contain in general news or web crawled texts that don’t mention concrete actions and concrete objects so often, we choose to build a smaller but coherent corpus of do-it-yourself instructions, with the assumption that it will contain frequent instances of concrete language.

We crawled from the website instructables.com all the titles and descriptions for the projects available online in six categories (e.g. technologies, workshop, living, food, play, outside). Cleaned of the html code, the instructables.com corpus has 17M tokens; each project was parsed with the Stanford parsed (de Marneffe and Manning 2008). To test how useful is a do-it-yourself instructions corpus with respect to a generic one, we represent each noun in the following experiment as a vector extracted from GoogleNews with word2vec toolkit (Mikolov et al. 2013) but also as a vector extracted from the instructables.com corpus trained with the same toolkit. These are the purely textual representations we experimented with; to complement this knowledge and to test the relevance of situational properties extracted through distributional models of semantics, we added extracted information about dimensions, weight and affording parts for 143 objects.

The list of objects’ parts that afford grasping and are component of the pattern extracted for the feature ”affording parts” has been derived with a psycholinguistic test (cf. De Felice 2015). Thirty students of the University of Pisa were interviewed

Table 2: Precision and recall for 8 combinations of features.

features	Precision	Recall
instructables.com	0.113	0.336
GoogleNews	0.113	0.336
weight	0.364	0.406
dimensions	0.413	0.517
dimensions+weight	0.561	0.531
affording parts	0.25	0.399
instructables.com + all	0.443	0.552
GoogleNews + all	0.458	0.559

and presented with 42 images of graspable entities. For each picture, they were asked to describe in the most detailed way how they would have grasped the object represented. Among the objects depicted, there were 31 artefacts. From the interviews recorded for these artefacts, we extracted all nouns denoting objects' parts that were named as possible target of the grasp (e.g. the handle for the bag, the cup or the ladle). The list of 78 nouns was then translated in English.

3.1 Classification Experiment

The experiment is based on a multi-label classification, since our dataset consists of 143 nouns denoting artifacts, annotated according to 6 categories. The implementation of Support Vector Multi-Classification is based on LibSVM software (Chang and Lin 2001) in WEKA with 10 fold cross-validation. Table 2 reports the results in terms of precision and recall. It is clear that the combination of all the features produces the best performance, with a slightly better result obtained when vectors trained on GoogleNews corpus are involved.

The overall performance is influenced by the fact that some classes are smaller in the gold standard, which constitutes the training and the test set, thanks to the 10 fold cross-validation. For this reason, we experimented with the same settings, but including just the 91 nouns that belong to the onehand or twohand classes. In Table 3, results show again that dimensions and dimensions plus weight produce good results, even if they do not improve the performance when combined with distributional vectors. Again, affording parts co-occurences produce the worst performance, mainly because the list of affording parts was originally derived for only 31 artefacts, and not for all the objects considered in our experiment.

Table 3: Precision and recall for 8 combinations of features on two classes dataset.

features	Precision	Recall
GoogleNews	0.846	0.846
weight	0.715	0.714
dimensions	0.851	0.846
dimensions+weight	0.831	0.802
affording parts	0.63	0.615
GoogleNews + all	0.846	0.846

4 Conclusions and Future Works

In this paper we test how distributional representations of nouns denoting artifacts can be combined with physical information about their dimensions and weights automatically extracted from an e-commerce website and with co-occurrence information about their affording parts as found in a corpus of do-it-yourself instructions. The starting hypothesis - concerning grasping possibilities as manipulative actions an object can be involved in - was that they are conceptually a combination of situational and physical properties.

As a consequence, we expect the best performance from a mixed features models. This hypothesis was in part confirmed; even if the overall performance is not good enough for the implementation of a module that automatically classifies grasping possibilities for objects in embodied robotics, it is evident that only if we combine different knowledge sources about the physical world we can improve conceptual classification about concrete objects.

These results are in line with the current trend to mix textual and visual features from computer vision algorithms (Bruni et al. 2012) in order to go beyond the limitations of purely textual semantic representations that cannot encode information about colors, dimensions, shapes etc. As future work we plan to integrate the features used for the experiment in this paper with representations of words as bag of visual words derived from the scale-invariant feature transform (SIFT) (Lowe 1999) algorithm that in computer vision helps to detect and describe local features in images.

References

Barsalou, L.W. 2002. Being there conceptually: simulating categories in preparation for situated action. *Representation, Memory, and Development: Essays in Honor of Jean Mandler*,1–15.

Baroni, M., Dinu, G. and Kruszewski, G. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors Proceedings of ACL 2014 (52nd Annual Meeting of the Association for Computational Linguistics), East Stroudsburg PA: ACL, 238-247.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 2012. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. *Proceedings of ACM Multimedia*, 12191228.

Chaigneau, S.E., Barsalou, L.W., and Sloman, S. 2004. Assessing the causal structure of function. *Journal of Experimental Psychology: General*, 133: 601-625.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:127:27. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Katrin Erk. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. Language and Linguistics Compass, 6(10):635653.

De Felice, I. (in press). Objects parts afford action: evidence from an action description task. In V. Torrens (ed.), Language Processing and Disorders. Newcastle: Cambridge Scholars Publishing.

De Felice, I. (2015). Language and Affordances. PhD thesis, University of Pisa, Italy.

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston: Houghton Mifflin.

Lowe, D.G. 1999. Object recognition from local scale-invariant features. *International Conference on Computer Vision*, pp. 1150-1157.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.

Russo, I., De Felice, I., Frontini, F., Khan, F., and Monachini, M. (2013). (Fore)seeing actions in objects. Acquiring distinctive affordances from language. In B. Sharp, and M. Zock (eds.), Proceedings of The 10th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2013 (Marseille, France, 15-17/10/2013), 151-161.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: vector space models of semantics. J. Artif. Int. Res., 37(1):141188, January.