Grounding the Lexical Sets of Causative-Inchoative Verbs with Word Embedding

Edoardo Maria Ponti
University of Cambridge
ep490@cam.ac.uk

Elisabetta Jezek
Università degli Studi di Pavia
jezek@unipv.it

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Abstract

English. Lexical sets contain the words filling the argument positions of a verb in one of its senses. They can be extracted from corpora automatically. The purpose of this paper is demonstrating that their vector representation based on word embedding provides insights onto many linguistic phenomena, such as causative-inchoative verbs. A first experiment aims at investigating the internal structure of the sets, which are known to be radial and continuous categories cognitively. A second experiment shows that the distance between the intransitive subject set and transitive object set is correlated with the spontaneity of the event expressed by the verb, defined according to morphological coding and frequency.

Italiano. *I set lessicali contengono le parole che occupano le posizioni argomentali di un verbo in una delle sue accezioni, e possono essere estratti in modo automatico dai corpora. L'obiettivo di questo articolo è dimostrare che la loro rappresentazione vettoriale illumina alcuni fenomeni linguistici, come i verbi ad alternanza causativo-incoativa. Un esperimento investiga la struttura interna degli insiemi, che a livello cognitivo sono ritenuti categorie radiali e continue. Inoltre, un secondo esperimento mostra che la distanza fra l'insieme dei soggetti intransitivi e l'insieme degli oggetti transitivi è correlata alla spontaneità dell'evento espresso dal verbo, definita secondo la marca morfologica e la frequenza.*

1 Introduction

Lexicographic attempts to cope with verb sense disambiguation often rely on "lexical sets" (Hanks, 1996), which represent the lists of corpus-derived words that appear as arguments for each distinct verb sense. The arguments are the "slots" that have to be filled to satisfy the valency of a verb (subject, object, etc.). For example, {*gun*, *bullet*, *shot*, *projectile*, *rifle*...} is the lexical set of the object for the sense 'to shoot' of *to fire*. In previous works, e.g. Montemagni et al. (1995), lexical sets were collected manually and were compared through set analysis. The measure of similarity between two sets was proportional to the extent of their intersection. We believe that possible improvements may stem from deriving the lexical sets automatically and from exploiting the semantic information of the fillers fully. In this work, we devise an extraction method from a huge corpus and use a distributional semantics approach to perform our analyses. More specifically, we represent fillers as word vectors and compare them with spatial distance measures. In order to test the relevance for linguistic theory of this approach, we focus on a case study, namely the properties of verbs undergoing the causative-inchoative alternation. Section 1.1. outlines a framework for word embeddings and section 1.2 introduces the case study. Section 2 presents the method and the data, whereas section 3 reports the results of a couple of experiments.

1.1 Word Embedding

The full exploitation of the semantic information inherent to argument fillers for verbs can take advantage from some recent developments in distributional semantics. Recently, efficient algorithms have been devised mapping each word of a vocab-

ulary into a corresponding vector of n real numbers, which can be thought as a sequence of coordinates in a n-dimensional space (Mikolov et al., 2013). This mapping is yielded by unsupervised machine learning, based on the assumption that the meaning of a word can be inferred by its context, i.e. its neighbouring words in texts. This model has some relevant properties: the geometric closeness of two vectors corresponds to the similarity in meaning of the corresponding words. Moreover, its dimensions have possibly a semantic interpretation.

1.2 Causative-Inchoative Alternation

A possible testbed for the usefulness of representing the argument fillers as vectors are the verbs showing the so called causative-inchoative alternation. These verbs appear either as transitive or intransitive. In the first case, an agent brings about a change of state; in the second, the change of a patient is presented as spontaneous (e.g. *to break*, as in "Mary broke the key" vs. "the key broke").

The two alternative forms of these verbs can be morphologically asymmetrical: in this case, one has a derivative affix and the other does not. The first is labelled here as "marked", the second as "basic". Italian verbs with an asymmetrical alternation derive from the phenomenon of anti-causativization. The intransitive form is marked since it is sometimes preceded by the clitic *si* (Cennamo and Jezek, 2011). Haspelmath (1993) maintain that verbs that show a preference for a marked causative form (and a basic inchoative form) cross-linguistically denote a more "spontaneous" situation. Spontaneity is intended by the author as the likelihood of the occurrence of the event without the intervention of an agent. This work is non-committal with respect to whether spontaneity be an actual semantic factor. Rather, it is considered a notion useful for labelling the observed variations in morphology and frequency.

In this way, a correlation between the form and the meaning of these verbs was demonstrated. Moreover, Samardzic and Merlo (2012) and Haspelmath et al. (2014) argue that verbs that appear more frequently (intra- and cross-linguistically) in the inchoative form tend to morphologically derive the causative form, too. This time, the correlation holds between form and frequency. Vice versa, situations entailing agentive participation prefer to mark the inchoative form and occur more frequently in the causative form.

2 Previous Work

In the literature, many methods are available for the automatic detection of verb classes, such as causative-inchoative verbs. They exploit features based on argument alternations, such as subcategorization frames (Joanis et al., 2008). The identification of verb classes displaying a diathesis alternation was also performed through the analysis of selectional preferences. Most notably, the lexical items were compared via distributional semantics (McCarthy, 2000).

These features were usually induced from automatic parses of heterogeneous and wide corpora (Schulte Im Walde, 2000). In particular, the extraction of subcategorization frames was refined including e.g. noise filters based on frequency (Korhonen et al., 2000). Our work is inspired by these attempts to automatically induce lexical information regarding verbs, but its direction of research is reversed. Indeed, rather than classifying verb classes given this information, it analyses this information given a verb class in order to shed light on its properties from the perspective of linguistic theory.

3 Data and Method

The data are sourced from a sample of ItWac, a wide Italian corpus gathered through web crawling (Baroni et al., 2009). This sample was further enriched with morpho-syntactic information through the MATE-tools parser (Bohnet, 2010)[1] and filtered by sentence length (< 100). Eventually, sentences in the sample amounted to 2,029,454 items. A target group of 20 causative-inchoative verbs was taken from Haspelmath et al. (2014): they are listed here based on the reported transitive/intransitive frequency ratio, from the highest to the lowest.

close > open > improve > break > fill > gather > connect > split > stop > go out > rise > rock > burn > freeze > turn > dry > wake > melt > boil > sink

The extraction step consisted in identifying their argument fillers inside the sentences in the sample. In particular, the arguments considered were the subjects of intransitives (S) and objects

[1] LAS scores for the relevant dependency relations: 0.751 with dobj (direct object), 0.719 with nsubj (subject), 0.691 with nsubjpass (subject of a passive verb).

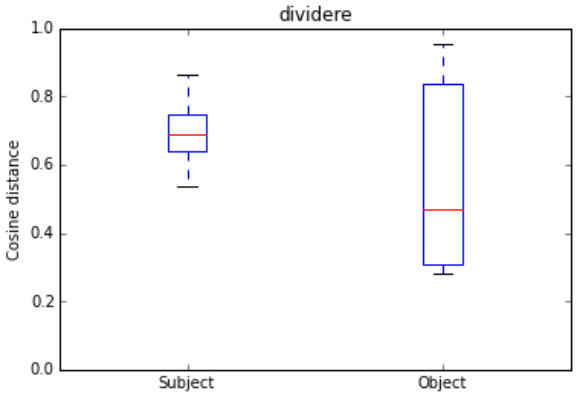

Figure 1: Distance of vectors from their centroid.

(O) (Dixon, 1994).[2] These arguments are relevant because they are deemed to share the same fillers (Pustejovsky, 1995).

These operations resulted in a database where each verb lemma had a single entry and was associated with a list of fillers, divided by argument type. With this procedure, lexical sets were extracted automatically, although they were not divided by verb sense. Afterwards, each of the argument fillers was mapped to a vector relying on a space model pre-trained through Word2Vec (Dinu et al., 2015).[3]

4 Experiments

In order to bring to light the linguistic information concealed in the automatically extracted lexical sets, we devised two experiments. One investigates the internal structure of lexical sets. In fact, previous works based on set theory treated them as categoric sets, of which a filler is either a member or not. Research in psychology, however, has long since demonstrated that the members of a linguistic set are found in a radial continuum where the most central one is the prototype for its category, and those at the periphery are less representative (Rosch, 1973; Lakoff, 1987).[4] Word vectors allow to capture this spatial continuum.

Once the fillers have been mapped to their respective vectors, a lexical set appears as a group of points in a multi-dimensional model. The centre of this group is the Euclidean mean among the vectors, which is a vector itself and is called centroid. In the first experiment, we calculated the coordinates of the centroid of the lexical sets S and O for any selected verb[5]. Then we evaluated the cosine similarity of every vector member of the sets from its centroid. The value of this metric goes from 0 (overlap) to 1 (maximum distance) and is useful to evaluate how far a filler is from its prototype. We obtained two sets of cosine similarity values for each verb: these can be plotted as boxes and whiskers, like in Figure 1. The example represents those of *dividere* 'to split'. The rectangles stand for the values in the second and third quartiles, whereas the horizontal line for the median[6]. From all these distance values, we picked the median value for each lexical set. The plot of these medians for the S set and the O set of each verb ordered according to Haspelmath's ranking is shown in Figure 2.

Two main results can be observed from these plots: the S lexical set lies in a more compact range of distances, whereas O is more scattered. On the other hand, the vectors of S tend farther from the centroid. This is demonstrated by the ranges where their distance values fall. Moreover, the averages of medians for the ten verbs on the left part of the scale (frequently transitive) and for the ten verbs on the right (frequently intransitive) were compared. The average median in S was 0.696567 for the former and 0.585263 for the latter. The average median in O was 0.556878 for the former and 0.522418 for the latter. This shows that the variation in O appears to be random. On the other hand, the median of the distances in S is normally lower for verbs that lie in the bottom half of the Haspelmath's scale.

The second experiment consisted in estimating the cosine distance between the centroid of S and the centroid of O for each verb. This operation was aimed at finding to which extent the lexical sets of S and O overlap. In fact, Montemagni et al. (1995) and McCarthy (2000) assessed in a corpus some asymmetries between these lexical sets, which in principle should share all their members.

[2] Subjects of forms with *si* were treated as intransitive subjects. Subjects of passive verbs were treated as objects.

[3] It was generated by a CBOW algorithm with negative sampling, 300 dimensions, a context window of 10 tokens, pruning of infrequent words and sub-sampling.

[4] For previous work on lexical sets considering prototypicality in the context of the notion of shimmering, see Jezek and Hanks (2010).

[5] Every filler was weighted proportionally to its absolute frequency.

[6] The median is the value separating the higher half of the ordered values from the lower half.

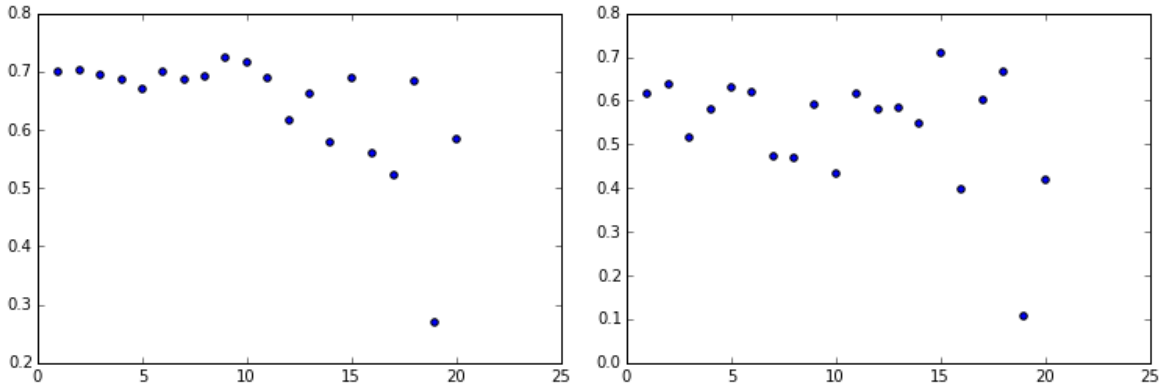

Figure 2: Medians of S (left) and O (right) distances for verbs ranked by position in Haspelmath's scale.

Inspecting our results, the distance between S and O seems to behave as a measure of spontaneity, intended as cross-linguistic frequency and morphological markedness of a verb: the more the centroids tend to be set apart, the more the verb tends to have a morphologically unmarked and more frequent intransitive form. In fact, we compared the ranking of 20 alternating verbs according to the ratio of their cross-linguistic frequency of transitive and intransitive forms (Haspelmath et al., 2014) and a ranking based on the centroid distances of the same verbs. Both these rankings are plotted in Figure 3: every verb is associated with its position in the two scales.

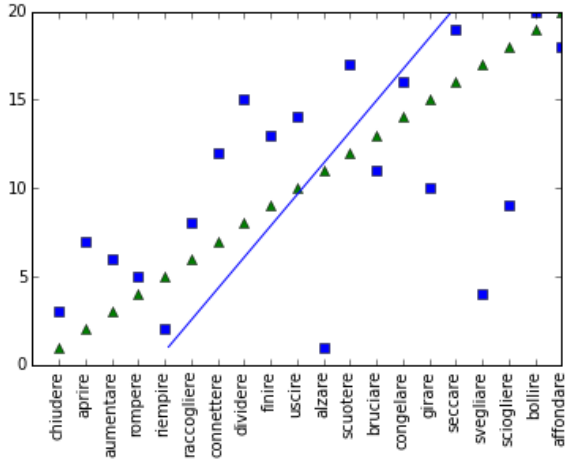

Figure 3: Ranking based on cross-linguistic form frequencies (green triangles) against ranking based on distance of the centroids of S and O in Italian (blue squares).

Both scales display a common tendency. In particular a Spearman's ranking test was performed over them, yielding a mild positive correlation of $\rho = 0.56391$ with a quite strong confidence, i.e. with $p < 0.01$.[7]

5 Discussion

The representation of lexical sets of Italian causative-inchoative verbs as vectors was demonstrated to provide insights into their internal structure and their relation with spontaneity defined according to morphological coding and frequency. The distances of the objects appeared to be distributed more uniformly, whereas those of the intransitive subjects more densely and remotely from the centroid. This difference cannot stem from the frequency of anaphoric fillers (contrary to transitive subjects), since both these argument positions share the discursive function of introducing new referents, and are hence occupied by fully referential fillers (Du Bois, 1985).

Moreover, the medians of the distances of the subject fillers from their centroid were shown to vary. An interpretation is that they are sensible to the frequency scale: this implies that frequently transitive (hence, non-spontaneous) verbs have semantically less homogeneous sets of referents, since they are farther from the prototype. Possibly this discovery can be related with the fact that non-spontaneous verbs impose less selectional restrictions on subjects (McKoon and Macfarland, 2000).

The lack of a perfect correlation between these vector distance and frequency measures is maybe due to errors in the automatic extraction and data sparseness for the former, or an insufficient sample

[7] An alternative measure was considered for the ranking: the cardinality of the S-O intersection weighted by the set union. In this case, Spearman correlation was $\rho = 0.42255$, but it was not significant because $p \approx 0.06$.

of languages in the typological survey of Haspelmath et al. (2014) for the latter. A possible interpretation of the correlation is that the entities capable of bringing about a change of state and those that undergo it are indiscernible only for non-spontaneous verbs. Studies on causer entities related them not only with the feature of agentivity, but also in general with the so-called 'teleological capability' (Higginbotham, 1997).

6 Conclusion

Our work provided evidence that lexical sets of Italian causative-inchoative verbs are continuous and radial categories, whose distribution around the prototype vary to a great extent. It is sensitive to the grammatical role and sometimes to the position of the verb in the so-called spontaneity scale. Moreover, a correlation was discovered between the distance between transitive object and intransitive subject lexical sets of a given verb and its cross-linguistic tendency to appear more frequently as intransitive or as transitive. Figure 4 is a synopsis of this result in the context of the correlations established in previous works.

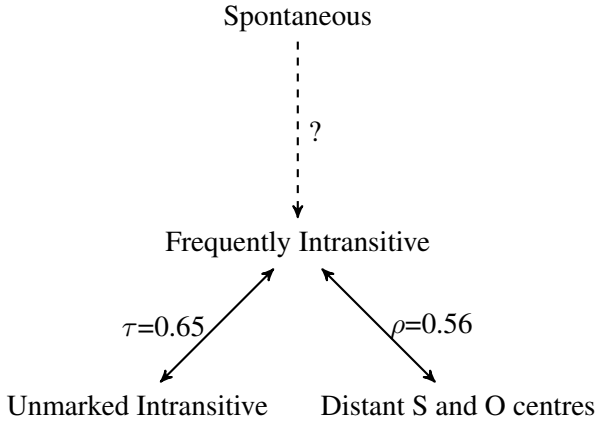

Figure 4: Synopsis of correlations among features of causative-inchoative verbs. The measures are based on Kendall Tau test (τ) and Spearman's ranking test (ρ).

In Figure 4, solid lines stand for correlations proven based on cross-linguistic evidence (frequency-form) and evidence from the Italian language (frequency-lexical sets). The dotted line, on the other hand, suggests the existence of and underlying motivation for the correlations, which nonetheless remains unproven and undetermined in its nature. Its possible validation is left to future research.

Future works should also choose different pretrained vector models, in order to try and replicate these results. In particular, the new vector models could be optimized for similarity through semantic lexica (Faruqui et al., 2015) or based on syntactic dependencies (Séaghdha, 2010). The experiments in this work may be extended to other languages, either individually or through a multi-lingual word embedding (Faruqui and Dyer, 2014).

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

Michela Cennamo and Elisabetta Jezek. 2011. The anticausative alternation in italian. *I luoghi della traduzione*, pages 809–823.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. *workshop contribution at ICLR 2015*.

Robert MW Dixon. 1994. *Ergativity*. Cambridge University Press.

John W Du Bois. 1985. Competing motivations. *Iconicity in syntax*, pages 343–365.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.

Patrick Hanks. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics*, 1(1):75–98.

Martin Haspelmath, Andreea Calude, Michael Spagnol, Heiko Narrog, and Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics*, 50(03):587–625.

Martin Haspelmath. 1993. More on the typology of inchoative/causative verb alternations. *Causatives and transitivity*, 23:87.

James Higginbotham. 1997. Location and causation. *Ms., University of Oxford*.

Elisabetta Jezek and Patrick Hanks. 2010. What lexical sets tell us about conceptual categories. *Lexis*, 4(7):22.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.

Anna Korhonen, Genevieve Gorrell, and Diana McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 199–206. Association for Computational Linguistics.

George Lakoff. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge University Press.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 256–263.

Gail McKoon and Talke Macfarland. 2000. Externally and internally caused change of state verbs. *Language*, pages 833–858.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

Simonetta Montemagni, Nilda Ruimy, and Vito Pirrelli. 1995. Ringing things which nobody can ring. a corpus-based study of the causative-inchoative alternation in italian. *Textus online only. 8 (1995), N. 2, 1995*, 8(2):1000–1020.

James Pustejovsky. 1995. *The generative lexicon*. The MIT Press.

Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology*, 4(3):328–350.

Tanja Samardzic and Paola Merlo. 2012. The meaning of lexical causatives in cross-linguistic variation. *Linguistic Issues in Language Technology*, 7(12):1–14.

Sabine Schulte Im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 747–753.

Diarmuid O Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444.