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

Edoardo Maria Ponti, Elisabetta Jezek, Bernardo Magnini

Università degli Studi di Pavia, Istituto Universitario di Studi Superiori, Fondazione Bruno Kessler
edoardomaria.ponti01@universitadipavia.it, jezek@unipv.it, magnini@fbk.eu

1. Introduction

Lexicographic attempts to cope with verb sense disambiguation often rely on “lexical sets” (Hanks, 1996), which represent the lists of corpus-derived argument fillers for each distinct verb sense. For example, {gun, bullet, shot, projectile, rifle…} is the lexical set for the sense ‘to shoot’ of to fire, while {employer, teacher, attorney, manager…} is the lexical set for sense ‘to dismiss’ of the same verb. In previous works, e.g. Montemagni et al. (1995), lexical sets have been collected manually and have been compared through set analysis. Their members (the fillers) are represented as strings of characters. The extent to which they are shared between two sets (i.e. the set intersection) is the measure of the set similarity. As a consequence, possible improvements may consist in deriving the lexical sets automatically and in exploiting the semantic information of the fillers fully. In this presentation, we devise an extraction method from a huge corpus and use a distributional semantics approach to achieve these improvements. 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 these innovations, we focus on a case study, namely the verbs undergoing the causative-inchoative alternation with respect to their spontaneity. Section 1.1. outlines a framework for word embeddings and section 1.2 introduces the case study. Section 2 presents the method, whereas section 3 reports the results of a couple of experiments based on it.

1.1. Word Embedding

The full exploitation of the semantic information inherent to argument fillers can take advantage from some recent developments in distributional semantics. Mikolov et al. (2013a) and Mikolov et al. (2013b) devised Word2Vec, an efficient algorithm mapping words into vectors of n real numbers, which can be thought as a sequence of coordinates in a n-dimensional space. The model that makes this mapping possible is yielded through unsupervised machine learning, based on the assumption that the meaning of a word can be inferred by its context, i.e. neighbouring words. 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. Case Study

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 identical (labile verbs), display different roots (suppletion) or be both morphologically derived through different affixes. In addition, the two forms can be morphologically asymmetrical: one has a derivative affix and the other does not. The first is labelled here as “marked”, the second as “basic”. (Haspelmath, 1993) maintain that verbs that cross-linguistically show a preference for a marked causative form (and a basic inchoative form) denote a more “spontaneous” situation. In this way, a correlation between the form and the meaning of these verbs is demonstrated. Moreover, Haspelmath et al. (2014) argues 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. Nonetheless, from these studies the relationship between frequency and spontaneity is not entirely clear (see Haspelmath et al. 2014, note 17)). We expect an analysis of lexical sets for these verbs carried out exploiting the methodology based on word embedding to shed light on our understanding of causation and spontaneity in language.

2. 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 crawl tools parser (Bohnet, 2010) and filtered by sentence length (< 100). Eventually, sentences in the sample amounted to 2,029,454 items. The next step consisted in identifying the verbs and their argument fillers inside these sentences. In particular, the
arguments considered were subjects of transitives (A), subjects of intransitives (S) and objects (O) (Dixon, 1994). These operations resulted in a database where each verb lemma had a single entry and every filler form occurring with it in the corpus was stored together, divided by argument type. With this procedure, an automatic extraction of lexical sets was accomplished, although lexical sets in this way are not divided by verb sense. Afterwards, each of the argument fillers was mapped to a vector relying on a word embedding pre-trained through Word2Vec (Dinu et al., 2015).

3. Experiments

![Figure 1: Median of the cosine distances of O fillers from the centroid of their lexical set.](image)

In order to bring to light the information concealed in the automatically extracted lexical sets, two experiments have been devised. One investigates the internal structure of lexical sets. In fact, previous works based on set theory treated them as categoric: a member is either a member of 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). Word vectors allow to capture this spatial continuum. Once the fillers have been mapped to their respective sets, a lexical set appears as a group of points in a multi-dimensional model. The centre of this group is just the Euclidean mean among the vectors, which is a vector itself and is called centroid. In the first experiment, we calculated the co-ordinates of the centroid of S and O for any given verb, and then the cosine similarity of every vector member of the set 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; from each of them we picked the median value, i.e. the element found in the middle of the cosine distances ordered by value. Two main results can be observed: the S lexical set is denser, whereas O is more scattered. This is demonstrated by the ranges where their distance values fall. Moreover, the median of the distances in S is normally lower for verbs that are highly spontaneous according to the Haspelmath’s scale. The averages of medians for the ten verbs on the left part of the scale (the most non-spontaneous) and for the ten verbs on the right (the most spontaneous) were compared. The average median in S was 0.696567 for the former and 0.585263 for the latter. The average median in S was 0.556878 for the former and 0.522418 for the latter. A plot of medians of the distances in S for every verb in the scale is plotted in Figure 1.

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 for the alternating verbs. In fact, these are assumed to share all their members, but Montemagni et al. (1995) cast some doubts over the precision of this assumption by assessing in a corpus some asymmetries between these lexical sets. Interestingly, inspecting our results the distance between S and O seems to behave as a measure of the spontaneity of an event: the more the centroids tend to be set apart, the more the situation denoted by the verb is spontaneous. We thus compared 20 alternating verbs ranked according to the direction of their morphological derivation (Haspelmath et al., 2014) and the same verbs ranked according to their centroid distance (from 0 to 1). Both are plotted in Figure 2: every verb is associated with its position in the two scales.

![Figure 2: Ranking based on morphological derivation (green triangles) against ranking based on distance of the centroids of S and O (blue squares).](image)

Both the 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 $p < 0.01$. To sum up, variations in the median distance of O fillers and in the distance of S and O centroids are connected with spontaneity.

---

2 Subjects of forms with the $si$-clitics were treated as intransitive subjects. Subjects of passive voices were treated as objects.

3 Its characteristics are a CBOW algorithm with negative sampling, 300 dimensions, a context window of 10 tokens, pruning of infrequent words and sub-sampling.

4 Every filler was weighted proportionally to its frequency.
References


