Distributional Analysis of Verbal Neologisms: Task Definition and Dataset Construction

Matteo Amore
University of Pavia / Pavia, Italy
CELI Language Technology / Turin, Italy
matteo.amore01@universitadipavia.it

Stephen McGregor
LATTICE - CNRS & École normale supérieure / Montrouge, France
Université Sorbonne nouvelle Paris 3 / Paris, France
semcgregor@hotmail.com

Elisabetta Jezek
University of Pavia / Pavia, Italy
Department of Humanities
jezek@unipv.it

1 Abstract

English In this paper we introduce the task of interpreting verbal neologism (VNeo) for the Italian language making use of a highly context-sensitive distributional semantic model (DSM). The task is commonly performed manually by lexicographers verifying the contexts in which the VNeo appear. Developing such a task is likely to be of use from a cognitive, social and linguistic perspective. In the following, we first outline the motivation for our study and our goal, then focus on the construction of the dataset and the definition of the task.

Italian *In questo contributo introduciamo un task di interpretazione dei neologismi verbali (Vneo) in italiano, utilizzando un modello di semantica distribuzionale altamente sensibile al contesto. Questa attività è comunemente svolta manualmente dai lessicografi, i quali verificano il contesto in cui il Vneo appare. Sviluppare questo tipo di task può rivelarsi utile da una prospettiva linguistica, cognitiva e sociale. Di seguito presenteremo inizialmente le motivazioni e gli scopi dell'analisi, concentrandoci poi sulla costruzione del dataset e sulla definizione del task.*

1 Introduction: motivation and goals

Studying neologisms can tell us several things. From a lexicographic point of view, neologisms can show trends that a language is following. In our opinion, they can also shed light on various aspects related to linguistic creativity; when speakers use new words (coined by themselves, or recently coined by someone else), they expect that the hearer can understand what they have just said.[1] Reversing the perspective, from the point of view of the hearers, when they encounter a word for the first time, they are generally capable of making hypotheses about the meaning of that word. The process of understanding unknown words involves the employment of previously acquired information. This knowledge can come from various sources: experience of the world, education, and contextual elements;[2] in this contribution we focus on linguistic contextual (namely co-occurrence) information.

For computational linguistics, neologisms raise some intriguing issues: automatic detection (especially for languages which do not separate written words with blank spaces); lemmatisation; POS tagging; semantic analysis; and so forth.

In this paper we present the task we have developed in order to interpret neologisms, using a context-sensitive DSM described by McGregor *et al.* (McGregor et al., 2015). This model was built to represent concepts in a spatial configuration, making use of a computational technique that creates conceptual subspaces. With the help of this DSM we intend to analyse the behaviour of a sub-group of neologisms, namely verbal neologisms (see Amore 2017 for more background).

Our goal is primarily linguistic. We intend to investigate the interpretation of VNeo, measuring the semantic salience of candidate synonyms by way of geometries indicated by an analysis of co-occurrence observations of VNeos. For instance, we expect that the VNeo *googlare* 'to google' and a verb like *cercare* 'to search' are geometrically related in a *subspace* specific to the conceptual context of the neologism.

[1] This is not the case of neologisms created for advertising, brand names or marketing purposes in general (Lehrer, 2003:380).

[2] All of these aspects are investigated, for example, in the field of Contextual Vocabulary Acquisition (Rapaport & Ehrlich, 2000).

The interpretation of neologisms presents two main challenges: a) analysing verbs using vectors built only upon co-occurrences (thus excluding argument structures) is notoriously a difficult task for DSM;[3] b) neologisms are, by definition, words whose frequency is (very) low, because their use is (still) not widespread. Thus, it represents a challenge for DSM models exactly because the vectors for most VNeo will rely upon few occurrences. In order to evaluate our results, we will compare them with the ones obtained using the Word2Vec model (Mikolov et al., 2013a), and with a gold standard consisting in human judgments on semantic relatedness (synonymy). The paper is structured as follows. In section 2 we introduce the DSM model that we employ in our task, and in section 3 we describe the construction of VNeo dataset and the problems we encountered. Finally, in section 4 we outline the task and present some preliminary thoughts on expected results.

2 Distributional Semantic Modelling

DSM is a technique for building up measurable, computationally tractable lexical semantic representations based on observations of the way that words co-occur with one another across large-scale corpora. This methodology is grounded in the *distributional hypothesis,* which maintains that words that are observed to have similar co-occurrence profiles are likely to be semantically related (Harris, 1954; Sahlgren, 2008). In general, a DSM consists of a high-dimensional vector space in which words correspond to vectors, and the geometric relationship between vectors is expected to indicate something about the semantic relationship between the associated words. The relationship most typically modelled is general semantic *relatedness,* as opposed to more precise indications of, for instance, *similarity* (Hill et al., 2015), but distributional semantic models have been effectively applied to tasks ranging from language modelling (Bengio, 2009) to metaphor classification (Gutiérrez et al., 2016) and the extrapolation of more fine-grained intensional correspondences between concepts (Derrac and Schockaert, 2015).

Standard DSM techniques present two problems for the task of interpreting neologisms. First, distributional representations are predicated on many observations of a word across a large-scale corpus: it is the plurality of context which gives these representations their semantic nuance. Second, the spaces generated by standard approaches like matrix factorisation and neural networks are abstract, in the sense that their dimensions are not interpretable; as such, typical distributional semantic models are not sensitive to the context specific way in which meaning arises in the course of language use. McGregor et al. (2015) have proposed a *context-sensitive* approach to distributional semantic modelling that seeks to overcome this second problem by using contextual information to project semantic representations into lower dimensional conceptual perspectives in an on-line way.

This methodology entails the selection of sets of dimensions from a base space of co-occurrence statistics that are in some sense conceptually *salient* to the context being modelled. The selection of salient features facilitates the projection of subspaces in which the geometric situation of and relationship between word-vectors are expected to map to a specific conceptual context. This technique has been applied to tasks involving context sensitive semantic phenomena such as metaphor rating (Agres et al., 2016), analogy completion (McGregor et al., 2016), and the classification of semantic type coercion (McGregor et al., 2017).

With regard to the first problem of data sparsity, we propose that the facility of the dynamically contextual approach for handling the *ad hoc* emergence of concepts (Barsalou, 1993) should provide a way of mapping from relatively few observations of neologisms, possibly taken outside the data used to build the underlying model, to context specific perspectives on distributional semantic representations.

3 Verbal Neologisms: dataset, corpus and lemmatisation

We will now explain the methodology we use in our analysis, and describe the resources we exploit highlighting their main features.

3.1 Sources for the neologisms list

To select the VNeo to be analysed, we extract data from pre-existing lists of Italian neologisms. These lists come from three websites: a)

[3] Cf. Bundell et al., 2017 and Chersoni et al., 2016.

treccani.it[4] b) iliesi.cnr.it/ONLI/[5] c) accademiadellacrusca.it.[6] (a) and (b) are manually compiled and validated: they contain words manually found in some widely read newspapers but not (yet) included in Italian dictionaries, coherently with the lexicographical definition of neologisms (cf. Adamo & Della Valle 2017). (c) consists of a list of words that, according to the users of the website, should be included in dictionaries. There is no curating of these suggestions (except the removal of swearwords); thus some neologisms might already be included in dictionaries. We chose to use this list because it allows analysing words which are perceived as new from a community of Italian speakers. In this way we intend to highlight the perspective of the hearers encountering new words.

Within the lists, we select only the verbs, obtaining a set of 504 VNeo. Of these VNeo, we check their presence in the itTenTen16 corpus, which we will also use to create the distributional vector space. 340 VNeo are attested in the corpus: 108 have between 10 and 99 occurrences; 79 between 100 and 999 occurrences; and 26 have more than 1000 occurrences.

Instead of using heuristic techniques that might have identified neologisms within the corpus (e.g. computing less frequent words and manually checking their presence in dictionaries),[7] we chose to rely on lists because we intend to study words whose use is wider and not restricted only to the web domain.

3.3 itTenTen16 corpus

We conduct an analysis of the itTenTen16 corpus (Jakubíček et al. 2013) because it is the most up-to-date corpus available for Italian. It is also a web-based corpus, and so particularly well fitted to examine neologisms: in fact, the web and IT domain is a notable source of new words and, especially, of new loanwords. As the corpus dimensions are sizeable (4.9 billion tokens), we will use a random sample of the full corpus for purposes of computability. This sample will correspond to ⅕ of the original corpus.

Starting from the corpus, the base DSM is built based on observations of the most frequent 200,000 words (defined as *vocabulary*) and their contextual information, considering a co-occurrence window of 5 words on either side of a target word. For the purposes of this study, we consider the VNeos included in the vocabulary. In this way we obtain the base space.

In order to project a subspace contextualised by a VNeo, we consider the co-occurrence features with the highest mutual information statistics associate with that particular VNeo. So, for instance, we find the following salient features:

customizzare 'to customise' [city; modellazione; illustrato; type; batch; editare; nastro; segmentare; preferenza; iconico; ...]

resettare 'to reset' [reset; password; formattare; bios; clempad; clementoni; fonera; resettare; centralina; router; ...]

googlare 'to google' [telespettatore; pdf; tecnologia; informazione; addirittura; vi; chiave; invito; risposta; sapere; ...].

These features are associated with the maximum mutual information values in terms of their co-occurrence with each of the corresponding input neologisms.

Some other VNeos represented in the vocabulary are: *postare* ‘to post’, *taggare* ‘to tag’, *twittare* ‘to tweet’, *spammare* ‘to spam’, *attenzionare* ‘to warn’, *spoilerare* ‘share information that reveals plot of a book or film’, *bloggare* ‘to blog’, *loggare* ‘to log’, *switchare* ‘to switch’.

It is worth noting that we create vectors starting from lemmas (not tokens). Our analysis highlighted the presence of some inaccuracies in the automatic lemmatisation of neologisms,[8] which was already present in the original corpus.[9] In a future investigation we are planning to compare the results produced with the original lemmatised corpus against the results obtained from a corpus version, where the lemmatisation will be corrected. This correction process might be performed using regular expressions, in order

[4] http://www.treccani.it/magazine/lingua_italiana/neologismi (last consulted 10/04/2018)

[5] http://www.iliesi.cnr.it/ONLI/BD.php (last consulted 02/05/2018)

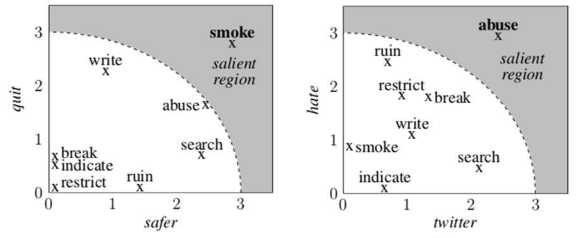
[6] http://www.accademiadellacrusca.it/it/lingua-italiana/parole-nuove (last consulted 02/05/2018)

[7] We are aware that this might correspond to the loss of some other neologisms contained in the corpus.

[8] Neologisms are not stored in common word-lists, and they are (usually) rare words, thus presenting difficulties for machine learning techniques.

[9] The lemmatisation is obtained using the TreeTagger tool (Schmid, 1994) with Baroni’s parameter file (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)

to capture specific VNeos token.[10]



(a) vape

(b) troll

Figure 1: Two subspaces projected based on two co-occurrence dimensions closely associated with the words (a) *vaped* and *vaping*, and (b) *trolled* and *trolling*, as observed in a small set of recent posts on Twitter. Among vectors for a number of candidate interpretations of neologisms, we see appropriate interpretations emerging based on distance from the origin in each contextualised subspace, based on PMI statistics extrapolated from co-occurrences observed across English language Wikipedia.

4. Interpreting VNeo using geometrical subspaces

As referenced in §1, our goal is to verify whether the meaning of a neologism can be induced from its context through distributional techniques, in particular by discovering verbs with salient geometric features in a contextualised subspace.

To this end, we organize the task as follows. Starting from a subset of the most frequent VNeos found in the corpus (§3), we first build subspaces for VNeos using the DSM model presented in §2. Subspaces are created by selecting the sets of dimensions that are conceptually salient to the context being modelled: each dimension in a subspace corresponds to a specific co-occurrence feature (i.e. a word). By finding a whole set of co-occurrences and using these to generate a relatively high-dimensional projection, we hope to establish a general contextualised conceptual profile and to overcome the peculiarities associated with low-frequency targets. For example, if the model finds that *googlare* 'to google' co-occurs with words like *nome* 'name', *indirizzo* 'address', and *sito* 'website', we use those co-occurrences as a basis for a projection of a *subspace* in which one could predict to find terms like *cercare* 'search' using geometric techniques.

Context can be defined in an open ended way in these models. For instance, the salient co-occurrence features of a single word can be used to generate a subspace. Small sets of words, either components of observed compositions (McGregor et al., 2017) or groups of conceptually related terms (McGregor et al., 2015) have also been used to generate semantically productive subspaces. In the small example illustrated in Figure 1, on the other hand, dimensions are defined explicitly in terms of the salient words associated with a small number of very recent observations of two different neologisms in use, specifically extrapolated from the salient co-occurrence features of Twitter posts in which the targeted neologisms are mentioned.

Contextualised subspaces can be explored in terms of the geometric features of word-vectors projected into those subspaces. So, for instance, McGregor et al. (2015) propose a norm method, by which word-vectors salient in a particular context will emerge as being far from the origin. This phenomenon is observed with appropriate interpretations percolating into the salient regions even in the low-dimensional toy examples illustrated in Figure 1, which involves a dynamically contextual DSM built from English language Wikipedia. Choices about context selection techniques, geometric characteristics of subspaces to be explored, and modelling parameters including dimensionality of projections will be the subject of our forthcoming experiments.

In order to evaluate the model, we will compare our results against the results obtained applying the Word2Vec model to the same corpus (Mikolov et al., 2013a).

With further investigations we will also test this model using a gold standard consisting of human judgments on VNeos interpretations collected for this purpose. Similarity judgments will be provided by two native speakers with significant background in linguistics. Specifically, the dataset will consist of verb pairs in which VNeo are grouped with more common verbs (*googlare* and *cercare*) based on human ratings collected in the form of a TOEFL-like multiple-choice synonymy test.[11]

[10] Regular expressions might be useful, within the corpus, to find an inflected form of a verb (lemmatised as it is) and replace it with the correct lemma: e.g. find lemma `googlav.` (meaning *googlavo, googlavi,* etc.*)* and replace it with *googlare*.

[11] Here the task is to determine, for a number of target words, the closest synonym from a choice of four alternatives.

4 Conclusion

The aim of the task presented here is to investigate the importance of linguistic context for the interpretation of neologisms, grounding the analysis in a context-sensitive DSM. With this task we intend to tackle issues connected with creativity processes and the environmental (contextual) sensibility typical of human cognition. In addition, we apply, for the first time, this DSM to Italian, providing a new semantic resource for the analysis of the language. Further studies may compare our results with other DSMs, and/or study what the semantic relations found with this specific approach reveal about other phenomena belonging to different linguistic levels (e.g. syntax).

References

Giovanni Adamo and Valeria Della Valle. 2017. *Che cos'è un neologismo?*. Carocci Editore, Roma.

Kat Agres, Stephen McGregor, Karolina Rataj, Matthew Purver, and Geraint A. Wiggins. 2016. Modeling metaphor perception with distributional semantics vector space models. In *Workshop on Computational Creativity, Concept Invention, and General Intelligence*, 08/2016.

Matteo Amore. 2017. I Verbi Neologici nell'Italiano del Web: Comportamento Sintattico e Selezione dell'Ausiliare. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, December 11-13, 2017.

Lawrence W. Barsalou. 1993. Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A.C. Collins, S.E. Gathercole, and M.A. Conway, editors, *Theories of memory*, pages 29–101. Lawrence Erlbaum Associates, London.

Yoshue Bengio. 2009. Learning deep architecture for AI. *Machine Learning*, 2(1):1–127.

Benjamin Blundell, Mehrnoosh Sadrzadeh, Elisabetta Jezek. 2017. Experimental Results on Exploiting Predicate-Argument Structure for Verb Similarity in Distributional Semantics. In Clasp Papers in Computational Linguistics, vol. 1, pages 99-106.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, Chu-Ren Huang 2016. Representing Verbs with Rich Contexts: an Evaluation on Verb Similarity, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* Association for Computational Linguistics, pages 1967–1972.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2011. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis,* 36:345–384.

Joaquín Derrac and Steven Schockaert. 2015. Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence*, 228:66–94.

E. Darío Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin K. Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, 41(4):665–695.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The tenten corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127.

Adrienne Lehrer. 2003. Understanding trendy neologisms. *Italian Journal of Linguistics*, 15:369–382.

Stephen McGregor, Kat Agres, Matthew Purver, and Geraint Wiggins. 2015. From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*, 6(1):55–89.

Stephen McGregor, Matthew Purver, and Geraint Wiggins. 2016. Words, concepts, and the geometry of analogy. In *Proceedings of the Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science (SLPCS)*, pages 39–48.

Stephen McGregor, Elisabetta Jezek, Matthew Purver, and Geraint Wiggins. 2017. A geometric method for detecting semantic coercion. In *Proceedings of 12th International Workshop on Computational Semantics*.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop Papers*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. Cognitive Science 34:1388–1429.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of NAACL-HLT 2018,* pages 2227–2237.

William J. Rapaport and Karen Ehrlich. 2000. A computational theory of vocabulary acquisition. In Stuart Charles Shapiro and Lucja M. Iwánska, editors, *Natural language processing and knowledge representation: language for knowledge and knowledge for language*. MIT Press, Cambridge, MA.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.