A complex network approach to semantic spaces: How *meaning* organizes itself

Salvatore Citraro[1], Giulio Rossetti[2]

[1] University of Pisa, Italy
salvatorecitraro939@gmail.com
[2] ISTI-CNR, Pisa, Italy
giulio.rossetti@isti.cnr.it

Discussion Paper

Abstract. We propose a complex network approach to the emergence of *word meaning* through the analysis of semantic spaces: NLP techniques able to capture an aspect of meaning based on distributional semantic theories, so that words are linked to each other if they can be substituted in the same linguistic contexts, forming clusters representing semantic fields. This approach can be used to model a mental lexicon of word similarities: a graph $G = (N, L)$ where N are words connected by some type of semantic or associative property L. Networks extracted from a baseline neural language model are analyzed in terms of global properties: they are small world and the probability of degree distribution follows a truncated power law. Moreover, they throw in a strong degree assortativity, a peculiarity that introduces us to the problem of semantic field identification. We support the idea that semantic fields can be identified exploiting the topological information of networks. Several community discovery methods have been tested, identifying from time to time strict semantic fields as crisp communities, linguistic contexts as overlapping communities or meaning conveyed by single words as communities produced starting from a seed-set expansion.

1 Introduction

In this work we assume distributional semantic theories - modeled by semantic spaces[1] - in order to analyze the complex structure of word meaning: words appearing in the same linguistic contexts form clusters representing semantic fields. In semantic spaces words are represented as vectors, whereas similar ones are near in terms of a geometric distance: therefore, it can be possible - through a similarity function - modeling some type of semantic or associative relatedness between words. Moreover, if we represent vectors as nodes, we can connect the

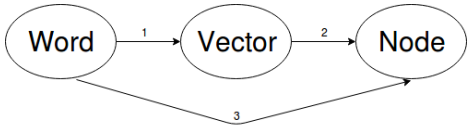

Fig. 1. (1) From texts to semantic spaces (2) From semantic spaces to networks (3) Can a graph have any linguistic reality?

highly similar ones with a link, letting a complex semantic network emerges. The scheme of the approach is summarized in Fig. 1.

The aim of the work is the identification of semantic fields through the exploitable topological information of networks. Treating semantic fields as communities (i.e., set of nodes tightly connected to each other), we want to partition a graph using several community discovery algorithms.

Details about the state-of-art of complex semantic networks will be given in Section 2. Data preparation (i.e., how complex networks have been extracted) will be described in Section 3. Global properties of networks will be analyzed in Section 4 - in terms of degree distribution, small world properties and assortativity. Section 5 will be about the analysis of the different types of semantic fields that we modeled. Section 6 will introduce several futures lines of research.

2 Complex semantic networks and the mental lexicon

The relationship between complex networks and language - in this case, its semantic level - is not trivial. The first metaphors of semantics as network of words refer to the models of semantic memory. Nowadays, they have been using to model a mental lexicon of word similarities, arguing how its structure may result in a complex system[11] as equal to biological, physical, social phenomenons, among others. Network Science offers a paradigm of research capable of determine the complexity of a system. Scale-freeness and small world properties are typical peculiarities of real complex systems: a network is scale-free if its degree distribution follows a power law and it is small world if clustering coefficient is higher and average path length is shorter than those of a random network. Practically, this means that, contrary to a random network, scale-free networks have a few number of highly connected nodes and a long tail of poorly connected ones. Starting from these measures, related works showed how complex semantic networks extracted from lexical databases, thesauri, association norms and treebanks annotated with the role arguments of verbs are scale-free and small world[7][8].

Complex semantic networks are graphs $G = (N, L)$ in which N are words connected by some type of semantic property L. Clarify the type of semantic property helps us to classify complex semantic networks from the aspect of meaning that they are modeling. Practically, in the previous examples nodes are *real words* and semantic properties represent relations like synonymy, hyponymy,

hypernymy, free associations or the arguments of verbs, while in network extracted from semantic spaces nodes are vectors connected by an high value of a similarity function. Literature points out how it is hard to verify if networks extracted from semantic spaces are scale-free, although the presence of small world phenomenon: it is argued how degree distributions could follow an exponential distribution[7] as well as a truncated power law[9] or a lognormal one[10] rather than a pure power law. This may be a constraint due to the representational framework necessary for the semantic space construction. However, the literature has concentrated more on global properties than on the meso-scale structure of all this networks, whereas more sensible aspects of word meaning complexity can be captured.

3 Data preparation

Given a training corpus $C = \{w_1, w_2, ..., w_n\}$ as a sequence of tokens and a semantic space as a vector space V in R, the basis of V is the set of the types of the corpus, i.e., the unique occurrence of a token, resulting in a vocabulary $T = \{t_1, t_2, ..., t_n\}$. The dimension of V is $R^{|T|}$, where $|T|$ is the length of vocabulary. Words are first represented as one-hot vectors $v(t_i)$, where 1 is the position of t_i in T. Then, they are embedded in a reduced dense space of dimension R^k, with $k \ll |T|$. Word2vec[12] is a technique to create word embeddings. SkipGram with Negative Sampling is its baseline model, that we used. It predicts a context from a word. Input vectors are one-hot vectors, while output ones are a probability distribution normalized with a softmax function (i.e., a language model). The model creates two types of word embedding from two matrices, W_I of dimension $|T| \cdot k$ and W_0 of dimension $k \cdot |T|$, where each row of W_I defines the embedding of the word t_i, namely h, and each row of W_0 defines the embedding of words when they appear in context with h, namely u. The softmax functions converts u in a probability distribution:

$$y_i = logp(t_{i-2}, t_{i-1}, t_{i+1}, t_{i+2}|t_i) = \frac{exp(u_c)}{\sum_{j \in T} exp(u_j)}$$

where $t_{i-2}, t_{i-1}, t_{i+1}, t_{i+2}$ is the context of a word if the window length was 2, u_c is the probability of the context to maximize and u_j is the rest to minimize. Instead of compute the function on all u_j in the vocabulary, Negative Sampling method does it on a small sample $u_m, m \in E$, where E is sampled using a biased unigram distribution.

Although the complexity of the framework, it can be simply implemented in Python with Gensim[13], a library allowing to tune the value of some hyperparameters like the dimension k of the embedding space (*size*), the length of context (*window*), the number of training words sampled by their frequency in the text (*min_count*) and the dimension of E (*negative*). Corpora used are a light version of *Italian Wikipedia* and *Paisa'*[14], a collection of Italian texts from web, both about 250 million tokens. Lemmas of open class words are used for training (better of lexemes to avoid morphosyntactic relations). Tuned values

are $size = 200$, $window = 10$, $negative = 10$ for both corpora, $min\_count = 10$ for *Wikipedia* and $min\_count = 200$ for *Paisa'*. Obtained the word embeddings, the cosine similarity is used to quantify the distance between vectors:

$$sim(\vec{v_1}, \vec{v_2}) = \frac{\vec{v_1} \cdot \vec{v_2}}{\|\vec{v_1}\| \|\vec{v_2}\|}$$

Then, vectors are represented as nodes, with L_{max} as the cosine similarity distribution, namely the number of all distances between all vectors. It is needed a graph in which $L \ll L_{max}$, whereby L must contain strong semantic relations: thus, L is chosen with ϵ-*method*[7], connecting vectors if and only if their cosine similarity exceeds a threshold ϵ. Value of ϵ are 0.5 (only for global network analysis) and 0.65, due to the computational costs of community discovery tasks.

4 Network analysis

Global level is analyzed in terms of degree distribution, small world properties and assortativity by degree. Fig. 2 sums up the analyses on degree distribution and assortativity. As regards first one, it is applied the likelihood-ratio test (LR-test) (also visualized in Table 1) with *powerlaw* library[15]. The LR-test consists of the comparison of the goodness of fit of two models, i.e., how well one of them fits a set of observations: a truncated power law seems to be the better model.

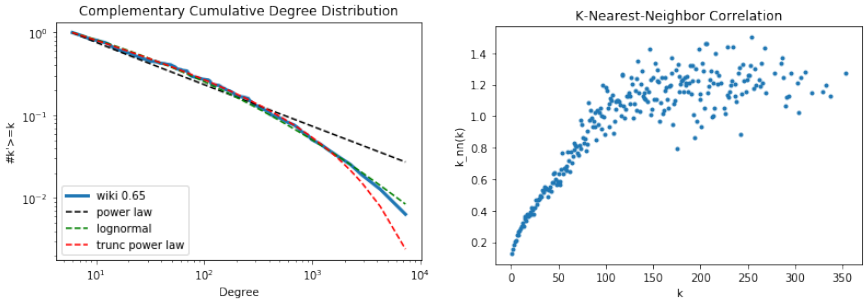

Fig. 2. Complementary cumulative degree distribution and degree assortativity by correlation between k and $k_{nn}(k)$ in *Wikipedia 0.65*

Power law, truncated power law, exponential and lognormal distributions have been compared. Assortativity describes if nodes, on average, connect to nodes with similar degree. Average K-Nearest-Neighbors $k_{nn}(k)$ is used for assortativity, i.e, the average k_{nn} for each node with degree k, where k_{nn} is the average neighbors degree of each node: there is correlation between k and $k_{nn}(k)$, thus networks are assortative. In Table 2 a general description of global network properties shows high global clustering coefficient C and short average path length $\langle d \rangle$, thus networks are small world.

	PL vs. TPL		PL vs. Exp		PL vs. lN		TPL vs. Exp		TPL vs. lN	
Dataset	R	p	R	p	R	p	R	p	R	p
Wikipedia 0.5	-2.54	0.00	4.7	0.00	-1.8	0.00	5.1	0.00	1.09	0.27
Wikipedia 0.65	-1.46	0.04	1.03	0.30	-1.1	0.31	1.95	0.05	1.99	0.04
Paisa' 0.5	-4.96	0.00	8.8	0.00	-3.67	0.00	9.74	0.00	3.0	0.00
Paisa' 0.65	-3.5	0.00	-2.56	0.01	-2.24	0.02	-0.75	0.45	0.07	0.03

Table 1. LR-test (*PL* is for *Power Law*, *TPL* is for *Truncated Power Law*, *Exp* is for *Exponential* and *lN* is for *Lognormal*): if R is positive, the data is more likely in the first distribution; vice versa, if it is negative, and p is the significance value ($p < 0.05$).

Dataset	N	L	$\langle k \rangle$	C	$\langle d \rangle$	d_{max}
Wikipedia 0.5	55244	1702789	61.6	0.39	4.98	13
Wikipedia 0.65	35728	270861	15.1	0.36	7.5	22
Paisa' 0.5	22622	578934	51.5	0.38	4.03	11
Paisa' 0.65	17090	115633	13.5	0.39	6.63	19

Table 2. N is the number of nodes, L is the number of links, $\langle k \rangle$ is the average degree, C is the global clustering coefficient, $\langle d \rangle$ is the average path length, d_{max} is the diameter.

Hubs (i.e, highest degree nodes) suggest that words belong to semantic fields, e.g., in *Wikipedia* networks hubs are words like *iperpiressia, infiammazione, ulcerativo, etc*, while in *Paisa'* ones they are words like *facciata, marmoreo, porticato, etc*. If degrees represent the extent of a semantic field, these words could belong to clusters whose lengths depend on the number of nodes tightly connected to each other:hubs suggest how there might be huge semantic fields lexically richer than others.

Starting from this simple interpretation of degrees, we can think about the lexical richness as a property of semantic fields able to be captured exploiting topological information. Together with other aspects and properties of semantic fields, this will be argument of the next section.

5 Community discovery

Community discovery is the task of the detection of highly connected nodes in complex network structures. In a complex semantic network meso-scale structure is formed by semantic fields.

The idea to represent a semantic fields as a community is so explained: (i) graph clusters are strict semantic fields identified by crisp communities in which a node belongs to one and only one community; (ii) semantic fields are represented by linguistic contexts identified by overlapping communities in which a node belongs to more than one community; (iii) clusters are local semantic fields conveyed by single words, identified by communities produced starting from a seed-set expansion.

Community discovery algorithms produce different types of community on the basis of the specific topological property they choose to detect: e.g., Louvain[2] maximizes a modularity function, Infomap[3] utilizes the map equation framework, Label Propagation[4] uses network structure alone, Demon[5] starts

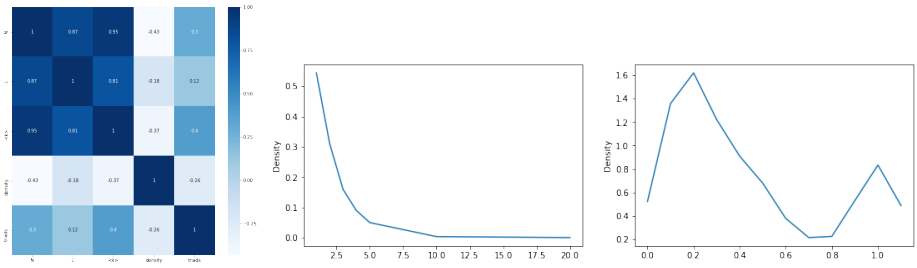

Fig. 3. From left to right (network: *Wikipedia 0.65*), (1) heatmap of correlation of Infomap partition, (2) overlapping distribution of Demon partition, (3) NF1 distribution of comparison between Lemon+Label Propagation and Ground Truth Partition

applying Label Propagation to ego-networks of nodes to merge them in a meso-scale structure, Lemon[6] is based on a seed-set expansion.

As regards crisp communities, Louvain and Infomap got good results. Partitions have been analyzed in terms of dimension N, number of edges L, internal edge density, average degree $\langle k \rangle$ and ratio of triads: Fig. 3, on the left, shows strong correlations between N, L and $\langle k \rangle$, both in Louvain and Infomap partitions. Both algorithms divide the graphs in consistent semantic fields: Louvain captures more general semantic domains while Infomap enhances them with granular partitions, e.g., if the largest community extracted with Louvain from *Wikipedia 0.65* belongs to the field of chemistry, Infomap can break up it in the subdisciplines of chemistry (organic chemistry, biochemistry, pharmacology, etc), consistently with the dataset from which knowledge is extracted. However, limits of crisp partitions are not trivial: the approach is not word-oriented, e.g. *molecola* or *atomo* (in according to graph partitioning, they belong to the chemistry domain) can appear in one and only one community, contrary to their concrete use in more than one domain. We focused on other algorithms capable of produce overlapping communities, in particular Demon. Partitions have been analyzed in terms of overlapping distribution, showed in Fig. 3, in the center. In some cases, the idea of linguistic context as semantic field can be captured. The term *Siria* is one of the nodes with the largest number of communities in a Demon partition with $\tau = 0.5$. A human analysis can easily interprets the communities in which the term is present as two different linguistic context: in the first one its semantics concerns with an *historical geographical entity*, in the second one with a *modern geographical entity*, e.g., a community is composed by terms like *anatolia, babilonese, egitto, eufrate, mesopotamia, persia, sumero*, another one by terms like *arabia, armenia, egitto, iran, iraq, marocco, turchia*. However, even this approach is not totally word-oriented.

In order to focusing on the semantics of single words, communities identified by a seed-set expansion are preferred. The proposed method chains Lemon and Label Propagation algorithms: starting from the seed-centered communities extracted by Lemon for each node, then Label Propagation is applied to break them into smaller and denser word sets. Advantages of this approach focus on

the possibility to capture polysemy, i.e., multiple meaning expressed by words. We take account of two terms, *stima* as example of polysemy and *pesca* as example of homonymy. As regards the first term, the partition could be coherent with its polysemic nature, ready to be interpreted both as *the price or value of a possession* (words in communities are *miliardo, dollaro, milione, sterlina, euro*) and as *an approximate measurement* (words in communities are *grossomodo, pressapoco, incirca*), but it is surely mismatched a potential third community composed by words denoting the meaning of *stima* as *an appreciation to others*. As regards the second term, only words related to the sense of the activity of fishing are find, without words related to the sense of the fruit. Then, limits of this approach may be related to the corpora and to the language model themselves, i.e. to the missing textual information and to the bias of word2vec models to create a unique embedding for homonymic words. Moreover, it was performed a Ground Truth Testing to compare and evaluate quality of communities produced in this third approach. Ground Truth Partitions have been extracted from Wikipedia itself, using the disambiguation pages: each hyperlink is a community composed by terms present in the hyperlinked page whose frequency was higher than 1. Filtered networks were compared with NF1 measure, the normalized harmonic mean of Precision and Recall[16], whose distribution is showed in Fig. 3, on the right. As expected, many of communities compared are dissimilar, while in those who get a perfect match the issue concerns the partial terms coverage due to the filtering (i.e., at most six or seven terms compared). Results were interesting but we need deeper quantitative evaluations to test the goodness of partitions: at the state-of-art - this is evident from the approach - there are not valid Ground Truth Partitions for these types of networks.

6 Conclusions

Word meaning has been modeled as a complex system self-organized in semantic fields. The notion of *word meaning* was strictly related to word vectors, i.e., computational representations of meaning. Global properties of networks extracted from word2vec models were consistent with the previous literature.Assortativity might be trace of a context-based representation of word meaning: hubs belongs to few giant semantic fields because they can be substituted in the same wide contexts. A question for future researches could be about interpretations of that: does the assortativity depend on a biased distribution of corpora or on the used language model? Or it reveal real hidden structures of semantic fields in texts? Semantic fields discovery was treated as a task of graph clustering instead of word vectors clustering, namely an alternative approach aiming to model several definitions of semantic field with several community discovery methodologies. We have obtained good results only exploiting the topological information. However, several lines of future research can be followed. Future approach could focus on community discovery methods able to integrate network topology and external information about nodes (e.g., attributes on their frequencies in texts or on the age in which they are learned or on their categories in other levels

of language analysis) to better represent a community structure in a complex semantic network.

Acknowledgment

This work is partially supported by the European Community's H2020 Program under the funding scheme "INFRAIA-1-2014-2015: Research Infrastructures" grant agreement 654024, http://www.sobigdata.eu, "SoBigData".

References

1. A. Lenci, *Distributional models of word meaning,* Annual Review of Linguistics, vol. 4, pp. 151171, 2018.
2. V. D. Blondel et al., *Fast unfolding of communities in large networks,* Journal of statistical mechanics: theory and experiment, 2008.
3. M. Rosvall and C. Bergstrom, *Maps of random walks on complex networks reveal community structure,* Proc Natl Acad SciUSA, vol. 105(4), p. 11181123, 2008.
4. U. N. Raghavan, R. Albert and S. Kumara, *Near linear time algorithm to detect community structures in large-scale networks,* Physical review E, vol. 76(3), 2007.
5. M. Coscia, G. Rossetti, F. Giannotti, D. Pedreschi, *Uncovering hierarchical and overlapping communities with a local-first approach,* ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 9(1), 2014
6. Yixuan Li et al. *Uncovering the small community structure in large networks: A local spectral approach,* Proceedings of the 24th international conference on world wide web. International World Wide Web Conferences Steering Committee, 2015.
7. M. Steyvers and J. B. Tenenbaum, *The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth,* Cognitive science, vol. 29, pp. 4178, 2005.
8. H. Liu, *Statistical properties of chinese semantic networks,* Chinese Science Bulletin, vol. 54, pp. 27812785, 2009.
9. A. Utsumi, *A complex network approach to distributional semantic models,* PLoS ONE, vol. 10(8), 2015.
10. I. Kajic and C. Eliasmith, *Evaluating the psychological plausibility of word2vec and glove distributional semantic models,* 2018.
11. S. De Deyne et al., *Large-scale network representations of semantics in the mental lexicon,* in M. N. Jones (Ed.), Frontiers of cognitive psychology. Big data in cognitive science, pp. 174-202, New York, NY, US: Routledge/Taylor & Francis Group, 2017.
12. T. Mikolov et al., *Distributed representations of words and phrases and their compositionality* Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, pp. 31113119, 2013.
13. R. Řehůřek and P. Sojka, *Software Framework for Topic Modelling with Large Corpora,* Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 45-50, 2010.
14. V. Lyding et al., *The PAISA' corpus of italian web texts,* Proceedings of the 9th Web as Corpus Workshop (WaC-9), pp. 3643, 2014.
15. J. Alsto, E. Bullmore, and D. Plenz, *powerlaw: A python package for analysis of heavy-tailed distributions,* PLoS ONE, 2014
16. G. Rossetti, L. Pappalardo, S. Rinzivillo, *A novel approach to evaluate algorithms detection internal on ground truth,* Complex Networks VII, 2016