Analysis of Italian Word Embeddings

Rocco Tripodi
Ca' Foscari University of Venice
`rocco.tripodi@unive.it`

Stefano Li Pira
University of Warwick
`stefano.li-pira@wbs.ac.uk`

Abstract

English. In this work we analyze the performances of two of the most used word embeddings algorithms, skip-gram and continuous bag of words on Italian language. These algorithms have many hyper-parameter that have to be carefully tuned in order to obtain accurate word representation in vectorial space. We provide an extensive analysis and an evaluation, showing what are the best configuration of parameters for specific analogy tasks.

Italiano. *In questo lavoro analizziamo le performances di due tra i più usati algoritmi di word embedding: skip-gram e continuous bag of words. Questi algoritmi hanno diversi iperparametri che devono essere impostati accuratamente per ottenere delle rappresentazioni accurate delle parole all'interno di spazi vettoriali. Presentiamo un'analisi accurata e una valutazione dei due algoritmi mostrando quali sono le configurazioni migliori di parametri su specifiche applicazioni.*

1 Introduction

The distributional hypothesis of language, set forth by Firth (1935) and Harris (1954), states that the meaning of a word can be inferred from the contexts in which it is used. Using the co-occurrence of words in a large corpus, we can observe for example that the contexts in which *client* is used are very similar to those in which *customer* occur, while less similar to those in which *waitress* or *retailer* occur. A wide range of algorithms have been developed to exploit these properties. Recently, one of the most widely used method in many natural language processing (NLP) tasks is word embeddings (Bengio et al., 2003; Mikolov et

al., 2010; Mikolov et al., 2013). It is based on neural network techniques and has demonstrated to capture semantic and syntactic properties of words taking as input raw texts without other sources of information. It represents each word as a vector such that words that appear in similar contexts are represented with similar vectors (Collobert and Weston, 2008; Mikolov et al., 2013). The dimensions of the word are not easily interpretable and, with respect to explicit representation, they do not correspond to specific concepts.

In Mikolov et al. (2013), the authors propose two different models that seek to maximize, respectively, the probability of a word given its context (Continuous bag-of-word model), and the probability of the surrounding words (before and after the current word) given the current word (Skip-gram model). In this work we seek to further explore the relationships by generating word embedding for over 40 different parameterizations of the continuous bag-of-words (CBOW) and the skip-gram (SG) architectures, since as shown in Levy et al. (2015) the choice of the hyper-parameters heavily affect the construction of the embedding spaces.

Specifically our contributions include:

- **Word embedding.** The analysis of how different hyper-parameters can achieve different accuracy levels in relation recovery tasks (Mikolov et al., 2013).

- **Morpho-syntactic and semantic analysis.** Word embeddings have demonstrated to capture semantic and syntactic properties, we compare two different objectives to recover relational similarities for semantic and morph-syntactical tasks.

- **Qualitative analysis.** We investigate problematic cases.

2 Related works

The interest that word embedding models have achieved in the NLP international community has recently been confirmed by the increasing number of studies that are adopting these algorithms in languages different from English. One of the first example is the Polyglot project that produced word embedding for 117 languages (Al-Rfou et al., 2013). They demonstrated the utility of word embedding, achieving, in a part of speech tagging task, performances competitive with the state-of-the art methods in English. Attardi et al. (2014) have done the first attempt to introduce word embedding in Italian obtaining similar results. They have shown that, using word embedding, they obtained one of the best accuracy levels in a named entity recognition task.

However, these optimistic results are not confirmed by more recent studies. Indeed the performance of word embedding are not directly comparable in the accuracy test to those obtained in the English language. For example, Attardi and Simi (2014) combining the word embeddings in a dependency parser have not observed improvements over a baseline system not using such features. Berardi et al. (2015) found a 47% accuracy on the Italian versus 60% accuracy on the English. The results may be a sign of a higher complexity of Italian with respect to English as we will see section 4.1.

Similarly, recent work that trained word embeddings on tweets have highlighted some criticalities. One of these aspects is how the morphology of a word is opaque to word embeddings. Indeed, the relatedness of the meaning of a lemma's different word forms, its different string representations, is not systematically encoded. This means that in morphologically rich languages with long-tailed frequency distributions, even some word embedding representations for word forms of common lemmata may become very poor (Kim et al., 2016).

For this reason, some recent contribution on Italian tweets have tried to capture these aspects. Tamburini (2016) trained SG on a set of 200 million tweets. He proposed a PoS-tagging system integrating neural representation models and a morphological analyzer, exhibiting a very good accuracy. Similarly, Stemle (2016) proposes a system that uses word embeddings and augment the WE representations with character-level represen-

HP	SG	CBOW
dim	200, 300, 400, 500	200, 300, 400, 500
w	3, 5	2, 5
m	1, 5	1, 5
n	1, 5, 10	1, 5, 15

Table 1: Hyper-parameters

tations of the word beginnings and endings.

We have observed that in these studies the authors used either the most common set-up of parameters gathered from the literature (Tamburini, 2016; Stemle, 2016; Berardi et al., 2015) or an arbitrary number (Attardi and Simi, 2014; Attardi et al., 2016). Despite the relevance given to these parameters in the literature (Goldberg, 2017) we have not seen studies that analyze the different strategies behind the possible parametrization. In the next section, we propose a model to deepen these aspects.

3 Italian word embeddings

Previous results on the word analogy tasks have been reported using vectors obtained with proprietary corpora (Berardi et al., 2015). To make the experiments reproducible, we trained our models on a dump of the Italian Wikipedia (dated 2017.05.01), from which we used only the body text of each articles. The obtained texts have been lowercased and filtered according to the corresponding parameter of each model. The corpus consists of 994.949 sentences that result in 470.400.914 tokens.

The hyper-parameters used to construct the different embeddings for the SG and the CBOW models are: the size of the vectors (dim), the window size of the words contexts (w), the minimum number word occurrences (m) and the number of negative samples (n). The values that these hyper-parameters can take are shown in Table 1.

4 Evaluation

The obtained embedding[1] spaces are evaluated on an *word analogy* task, using a enriched version of the Google word analogy test (Mikolov et al., 2013), translated in Italian by (Berardi et al., 2015). It contains 19.791 questions and covers 19 relations types. 6 of them are semantic and 13 morphosyntactic (see Table 2). The proportions of

---

[1]The trained vectors with the best performances are available at http://roccotripodi.com/ita-we

Morphosyntactic	Semantic
adjective-to-adverb	capital-common-countries
opposite	capital-world
comparative	currency
superlative (assoluto)	city-in-state
present-participle (gerundio)	regione capoluogo
nationality-adjective	
past-tense	
plural	
plural-verbs (3rd person)	
plural-verbs (1st person)	
remote-past-verbs (1st person)	
noun-masculine-feminine-singular	
noun-masculine-feminine-plural	
#10.876	#8.915

Table 2: Relation types

m=1	m=5	Berardi
3.227.282	847.355	733.392

Table 3: Vocabulary length

these two types of question is balanced as shown in Table 2.

To recover these relations two different methods are used: 3COSADD (Eq. 1) (Mikolov et al., 2013) and 3COSMUL (Eq. 2) (Levy et al., 2014) to compute vectors analogies:

$$3\text{CosAdd} \operatorname*{argmax}_{b^* \in V} cos(b^*, b - a + a^*) \quad (1)$$

$$3\text{CosMul} \operatorname*{argmax}_{b^* \in V} \frac{cos(b^*, b)cos(b^*, a^*)}{cos(b^*, a) + \epsilon} \quad (2)$$

These two measures try to capture different relations between word vectors. The idea behind these measures is to use the cosine similarity to recover the vector of the hidden word (b^*) that has to be the most similar vector given two positive and one negative word. In this way, it is possible to model relations such as *queen* is to *king* what *woman* is to *man*. In this case, the word *queen* (b^*) is represented by a vector that has to be similar to *king* (b) and *woman* (a^*) and different to *man* (a). The two analogy measures slightly differ in how they weight each aspect of the similarity relation. 3COSADD allows one sufficiently large term to dominate the expression (Levy et al., 2014), 3COSMUL achieves a better balance amplifying the small differences between terms and reducing the larger ones (Levy et al., 2014). As explained in Levy et al. (2014), we expect 3COSMUL to overperform 3COSADD in evaluating both the syntactic and the semantic tasks as it tries to normalize the strength of the relationships that the hidden term has both with the attractor terms and with the repellers term.

4.1 Experimental results

The results of our evaluation are presented in Figure 1. The main trend that it is possible to notice is that accuracy increases as the number of dimensions of the embedded vectors increases. This indicates that Italian language benefits of a rich representation that can account for its rich morphology. Another important trend that emerges is the fact that the parameters have the same effect on both algorithms and that they perform very differently on all the tasks. CBOW has very low accuracy compared to SG. We can also see that the dim hyper-parameter is not correlated with the dimension of the vocabulary (model complexity) as one should expect. In fact, with increasing values of dim the accuracy increases whatever is the value of m. This hyper-parameter heavily affects the vocabulary length (see Table 3). However the dim hyper-parameter seems to be correlated only with the accuracy in the semantic tasks while the performances on the morpho-syntactic tasks seems not to have a big bust increasing the dimensionality.

With respect to the size of the context (w) used to create the words representations we do not observe a clear difference between the 18 pairs both in the SG and in the CBOW. On the contrary a clear trend can be observed varying the n hyperparameter, with $n = 1$ the accuracy was significantly lower than the one we obtained with $n = 5$ or $n = 10$. Increasing the number of negative samples constantly increases the accuracy.

These results support also the claim put forward by (Levy et al., 2014) that the 3COSMUL method is more suited to recover analogy relations. In fact, we can see that on average the right bars of the plots are higher than the left.

4.2 Error analysis

If we restrict the error analysis to the most macroscopic differences in figure 1 we can compare three different parametrizations: SG-200 w5-m5-n1, SG-500 w5-m5-n1, SG-500 w5-m5-n10. In this way we can analyze the results obtained changing the number of dimensions of the vectors and the role played by n. In Table 4 the total num-
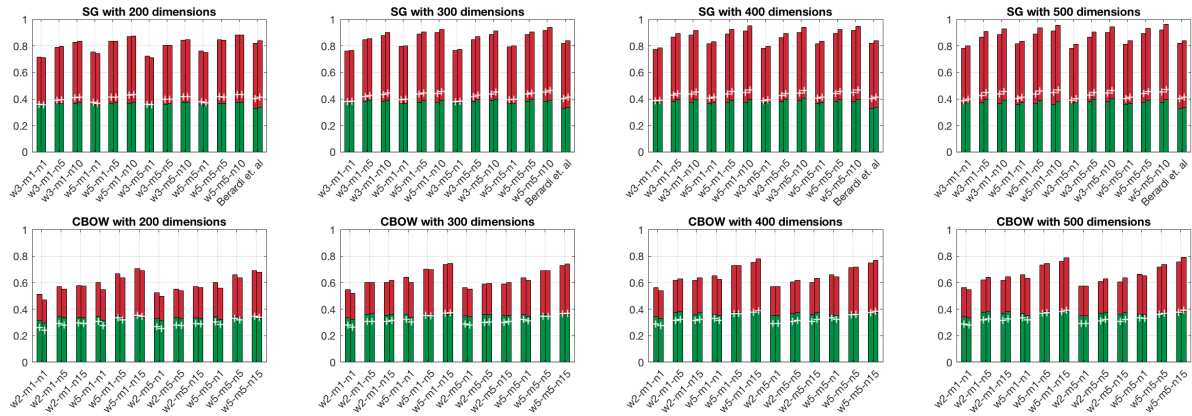
Figure 1: Results as accuracy with different hyper-parameters (y axis) using the 3COSADD (left bar) and the 3COSMUL (right bar) formula. The green part of the bars indicates the accuracy on the morphosyntactic task whereas the red one the accuracy on the semantic task. The $+$ sign on each bar indicates the accuracy on the entire dataset. The upper row of the figure shows the results of the SG algorithm and the bottom row the results of CBOW. The last two bars of the SG plots indicates the results obtained using the vectors made available by (Berardi et al., 2015)

Parametrization	#errors	#words
SG-200-w5-m5-n10	10.113	543
SG-500 w5-m5-n1	10.506	535
SG-500 w5-m5-n10	9.337	525

Table 4: Total number of errors and number of different words that have not been recovered

ber of errors and the number of different words that have not been recovered by each parametrization are presented. From this table we can see that most of the errors are done one a relatively small set of words. This phenomenon can be studied analyzing the most problematic cases. In Table 5 we can see the list of the most common errors ranked by frequency for each method. As we can

SG-200-w5-m5-n10	#	SG-500 w5-m5-n1	#	SG-500 w5-m5-n10	#
california	328	california	349	california	287
texas	223	texas	224	texas	165
arizona	164	arizona	164	arizona	145
florida	144	ohio	142	florida	124
ohio	135	florida	140	ohio	112

Table 5: Most common errors

see from these lists the errors are done on the same words and this because they are the most common in the dataset (e.g.: in the dataset there are 217 queries that require *Florida* as answer compared to the 55 of *Italia*). However if we compare the frequency of these errors in the analogy test within the three parametrisation we can observe an improvement of approximately 15% in accuracy with SG-500 w5-m5-n10. Indeed, despite many errors

are not recovered for any of the parametrisation, we can observe that approximately 21% of the errors are recovered under certain parametrizations (Table 6). To further investigate these improvements related to the aforementioned parametrisation we focused on one of the most frequent errors in the analogy test, the word *California*. As we can see from the list of the analogy test solved (Table 7) different parametrizations are helpful to solve different types of analogies. For example an increase in the dimensionality increases the accuracy, but mainly in analogy test with words that have a representation in the training data related to a wider set of contexts (*Houston*:*Texas*; *Chicago*:*Illinois*). The best parametrisation is obtained increasing the negative sampling. As we can see from the examples provided, the analogies are resolved thanks to a contextual similarity between the two pairs (*Huntsville*:*Alabama*; *Oakland*:*California*). In these cases the negative sampling could help to filter out from each representation those words that are not expected to be relevant for the words embeddings.

Similar types of improvement are noticed on analogy tests that contain a challenging word *predire* (*predict*). The results of this analysis are presented in Table 9 where it is possible to see that an higher dimensionality improves the accuracy of analogical tests containing open domain verbs (e.g.: *descrivere*, *vedere*). Similarly to the previous case, an higher dimensionality allows for fine

Parametrization	#errors solved
dim = 500 & n = 10	873
solo dim = 500	645
solo n = 10	927

Table 6: Solved errors

dim = 500 & n = 10	*solo* n = 10	*solo* dim = 500
Milwaukee Wisconsin Oakland California	Huntsville Alabama Oakland California	Houston Texas Oakland California
Shreveport Louisiana Oakland California	Baltimore Maryland Oakland California	Chicago Illinois Oakland California
Irvine California Shreveport Louisiana	Irvine California Phoenix Arizona	Denver Colorado Oakland California
Irvine California Baltimore Maryland	Arlington Texas Irvine California	Philadelphia Pennsylvania Oakland Calif
Sacramento California Henderson Nevada	Phoenix Arizona Sacramento California	Portland Oregon Oakland California
Sacramento California Orlando Florida	Huntsville Alabama Sacramento California	Tulsa Oklahoma Irvine California

Table 7: Examples of analogy tests solved.

grained partitions improving the correct associations between terms. However, also in in this case, the best parametrizations are obtained increasing the negative sampling or both the parameters. As we can see here both the present participle and the past tense pairs are correctly solved. These example provide a preliminary evidence of how negative sampling, filtering out non informative words from the relevant context of each word, is able to build representation by opposition that are beneficial both for semantic and syntactic associations.

Examples of words that almost always are not recovered correctly are presented in Table 10. A selected list of words problematic for all parametrizations is shown in Table 8. It contains plurals, feminine, currencies, superlatives and ambiguous words. The low performances on these cases can be explained by the poor coverage of these categories in the training data. In particular, it would be interesting to study the case of feminine and to analyze if it is due to a gender bias in the Italian Wikipedia, as a preliminary analysis of the most frequent errors that persist in all the parametrization seems to suggest. The words that have been benefited by the increase of n are:

ghana	slovenia	ucraino	portoghese
pakistan	giocando	zimbabwe	contessa
irlandese	serbia	namibia	messicano
migliorano		suonano	
scrivendo	implementano	maltese	giordania

the errors that have been introduced increasing this parameter are related to the words in Table 11. It is interesting to notice that given an error in an analogy test, it is possible to find the correct answer in the top five most similar words to the query. Precisely we observed this phenomenon in 26% of the cases for SG-200-w5-m5-n10, in 27% of the cases for SG-500-w5-m5-n1 and in 25% for SG-500-w5-m5-n1. Furthermore, approximately in 50% of these cases the correct answer is the second most similar. Most of the recovery errors are due to vocabulary issues. In fact, many words of the test set have no correspondence in the developed embedding spaces. This is due to the low frequency of

many words that are not in the training corpus or that have been removed from the vocabulary because of their (low) frequency. For this reason we kept the m hyper-parameter very low (e.g., 1 and 5), in counter-tendency with recent works that use larger corpora and then remove infrequent words setting m with high values (e.g., 50 or 100). In fact, with increasing value of m the number of not given answers increases rapidly. It passes from 300 ($m = 1$) to 893 ($m = 5$).

Some of the words that are not present in the vocabulary with $m = 1$ include plural verbs (1st person), that probably are not used by a typical Wikipedia editor and remote past verbs (1st person), a tense that in recent years is disappearing from written and spoken Italian. Some of these verbs are:

giochiamo	zappiamo	mescolai
affiliamo	implementai	
rallentiamo	rallentai	nuotai

In Berardi et al. (2015) the number of not given answer is 1.220. The accuracy of their embeddings, obtained using a larger corpus and using the hyper-parameters that perform well on English language, is always lower than those obtained with our setting, in both the morphosyntactic and the semantic tasks. This confirms that the regularization of the parameters is crucial for good representation of the embeddings, since the Berardi et al. (2015)'s model has been trained on a much larger corpus and for this should outperform ours. Furthermore, this model seems to have some tokenization problem.

5 Conclusions

We have tested two word representation methods: SG and CBOW training them only on a dump of the Italian Wikipedia. We compared the results of the two models using 12 combinations of hyper-parameters.

We have adopted a simple word analogy test to evaluate the generated word embeddings. The results have shown that increasing the number of

pilotesse	migliore	colori	meloni
pere	matrigna	figliastra	sua
real	lev	yen	mamma
kwanza	vantaggiosissimo	urlano	stimano
aquila	eroina	programmato	impossibilmente

Table 8: Always wrong

dim = 500 & *n* = 10	*solo* n = 10	*solo* dim = 500
dire detto predire predetto	cantare cantato predire predetto	descrivere descritto predire predetto
mescolare mescolando predire predicendo	correre correndo predire predetto	vedere visto predire predetto
predire predicendo generare generando	generare generando predire predetto	
rallentare rallentando predire predicendo	predire predicendo programmare programmando	
scoprire scoprendo predire predicendo	scrivere scrivendo predire predicendo	

Table 9: Examples of analogy tests solved.

SG-200-w5-m5-n10	#	SG-500 w5-m5-n1	#	SG-500 w5-m5-n10	#
capre	26	groenlandia	27	ratti	26
rapidamente	26	silenziosamente	27	ovviamente	25
dolcissimo	26	caldissimo	27	incredibilmente	25
apparentemente	26	occhi	27	grandissimo	25
andato	26	greco	27	malvolentieri	25

Table 10: Almost always wrong

irlanda	afghanistan	albania	egiziano
olandese	provvedono	francese	svizzero

Table 11: New errors

dimensions and the number of negative examples improve the performance of both the models.

These types of improvement seems to be beneficial only for the semantic relationships. On the contrary the syntactical relationship are negatively affected by the low frequency of many of its terms. This should be related to the morphological complexity of Italian. In the future it would be helpful to represent the spatial relationship regarding specific syntactical domain in order to evaluate the contribution of hyper-parametrization to syntactical relationship accuracy. Moreover future work will include the testing of these word embedding parametrizations in practical applications (e.g. analysis of patents'descriptions and books' corpora).

Acknowledgments

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *CoNLL-2013*, page 183.

Giuseppe Attardi and Maria Simi. 2014. Dependency parsing techniques for information extraction.

Giuseppe Attardi, Vittoria Cozza, and Daniele Sartiano. 2014. Adapting linguistic tools for the analysis of italian medical records.

Giuseppe Attardi, Daniele Sartiano, Chiara Alzetta, Federica Semplici, and Largo B Pontecorvo. 2016. Convolutional neural networks for sentiment analysis on italian tweets. In *CLiC-it/EVALITA*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Giacomo Berardi, Andrea Esuli, and Diego Marcheggiani. 2015. Word embeddings go to italy: A comparison of models and training datasets. In *IIR*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

John Rupert Firth. 1935. The technique of semantics. *Transactions of the philological society*, 34(1):36–73.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Zellig S Harris. 1954. Distributional structure. word, 10 (2-3): 146–162. reprinted in fodor, j. a and katz, jj (eds.), readings in the philosophy of language.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *AAAI*, pages 2741–2749.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Egon W Stemle. 2016. bot. zen@ evalita 2016-a minimally-deep learning pos-tagger (trained for italian tweets). In *CLiC-it/EVALITA*.

Fabio Tamburini. 2016. A bilstm-crf pos-tagger for italian tweets using morphological information. In *CLiC-it/EVALITA*.