Type of the Paper (Article)

What do we learn from word associations? Evaluating machine learning algorithms for the extraction of contextual word meaning in natural language processing

Epaminondas Kapetanios[1*], Saad Alshahrani[1], Anastasia Angelopoulou[1], Mark Baldwin[1]

[1] University of Westminster, School of Computer Science and Engineering; kapetae@westminster.ac.uk

* Correspondence: kapetae@westminster.ac.uk; Tel.: +44 20 79115000 ext. 64539

Abstract: "*You should know the words by the company they keep!*" has been one of the most famous slogans attributed to *John Rubert Firth*, 1957. This has ignited a whole school in linguistic research known as the British empiricist contextualism. Sixty years later, many un- or semi-supervised machine learning algorithms have been successfully designed and implemented aiming at extracting word meaning from within the context of a text corpus. These algorithms treat words, more or less, as vectors of real numbers representing frequencies of word occurrences within context and word meaning as positions of words in a high-dimensional vector space model. Word associations, in turn, are treated as calculated distances among them. With the rise of *Deep Learning (DL)* and other artificial neural networks based architectures, learning the positioning of words and extracting word associations as measured by their distances has further improved. In this paper, however, we revisited the main stream of algorithmic approaches and set the stage for a partly cross-disciplinary evaluation framework to judge about the nature of the extracted word associations by state-of-the-art machine learning algorithms. Our preliminary results are based on word associations extracted from the application of DL framework on a Google News text corpus, as well as on comparisons with human created word association lists such as word collocation dictionaries and psycholinguistic experiments. The results and conclusions provide some insights into the inherited limitations in interpreting the type of word associations and underpinning relations between words with inevitable consequences in other areas, such as extraction of knowledge graphs or image understanding.

Keywords: Machine Learning; Algorithms; Natural Language Processing, Deep Learning, Vector Space Models, Semantic Similarity, Distributional Semantics, Latent Semantic Analysis, Word2Vec

---

1. Introduction

There is a common belief that natural language processing (NLP) and understanding is theoretically a very complex process involving many different sources of information, particularly when this has to take place in real time. Natural language processing is concerned, to a great extent, with the automatic extraction of relations between words by means of statistical methods, usually measures of statistical co-occurrence. For this purpose, numerous un- or semi-supervised algorithms, e.g., *Latent Semantic Analysis* (LSA), *Latent Dirichlet Association* (LDA), have been introduced with the goal of extracting knowledge about relations between words. The foundations of these are co-occurrence statistics such as mutual information as well as comparison operators such as dice coefficient or Euclidean distance.

These computational approaches have different applications, for instance, Information Retrieval, disambiguation algorithms, speech recognition, or spellcheckers. They mostly utilize some sort of *Vector Space Models* (VSMs) as an attempt to represent the lexical meaning of words in terms of their

 doi:10.20944/preprints201805.0102.v2

positioning and distance from other words within a multi-dimensional space. This list of related approaches can be extended by neural network based architectures, as sparked by the recent success of Deep Learning (DL), which can be applied to improve learning of positions and associations between words within the underpinning vector space model. This space, in turn, provides a mechanism to measure the semantic similarity between words or between queries and document, as it is the case with Information Retrieval related tasks.

The historical motivation for computing relations between words, however, is attributed to John R. Firth [1], stating that meaning and context should be viewed as central in linguistics. Firth introduced the notion of collocation on the lexical level and defined it as the consistent co-occurrence of a word pair within a given context. "*You shall know a word by the company it keeps!*" is, perhaps, the most famous quotation attributed to Firth. The notion of collocation in its original meaning created the linguistic tradition and groundwork for the frequentist or empiricist tradition of British (corpus) linguistics. Apart from Firth, other representatives of the empiricist tradition have been Michael A. K. Halliday and John Sinclair. The central notion in their research, in extension to Firth, was that the empirical, even statistical, side of language use in text corpora could serve as a framework to describe and explain natural language. Indeed, many of the roots of the empirically motivated and statistical methodology in contemporary computational linguistics may be sought in this linguistic tradition. This can also be seen in various accounts on contemporary statistical NLP [2].

This frequentist corpus-based approach dedicated to an empirically grounded analysis of natural language, however, has been on the one side of a roughly dividing line of linguistic research.in the last half-century. On the other side, there is the *structural-lexicographic* approach which is mainly concerned with adequate representation forms of collocations within linguistic lexicons and dictionaries. The first dedicated and large-scale lexicographic study of collocations was undertaken for the English language by Benson et al. [3-5], which led to the publication of the BBI Combinatory Dictionary of English: A Guide to Word Combinations (in short: BBI) [3] outlines the motivation for a dictionary of word combinations and the kinds of information included in it.

The main goal has been to provide information on the general combinatorial possibilities of an entry word. Various types of combinatorial preferences are listed, such as e.g. whether there are any combinatorial preferences of verbs for nouns (e.g. "[to adopt, enact, apply] a regulation") or what the possible adverbial combinations (i.e. modifications) of a verb are (e.g. "to regret [deeply, very much]". There is also a distinction between grammatical and lexical collocations with the latter relying on part-of-speech patterns, such as verb-(preposition)-noun, adjective-noun or noun-noun, for permissible collocations in a natural language. For instance, "compose music" and "launch a missile" are permissible, while "compose a missile" is at least awkward.

At this point, it is worth noting the Meaning-Text Theory (MTT), which attempts to account for relations between lexical items in a language independent way. Within this framework, [6,7] attempt to come to terms with the idiosyncrasy of collocations by embedding them into a more semantically oriented layer of description. In the Meaning-Text Theory (MTT) lexical relations are used as a means of describing so-called institutionalized lexical relations. Based on MTT, a constant meaning linked to the combination between words is defined as a relation holding between two lexical items. These meanings and relations between lexical items are anchored as Lexical Functions (LFs) defined mostly on the semantic level.

Particularly, there are 36 syntagmatic LFs which are distinguished by their syntactic part of speech. Examples of LFs and their English realization are provided below:

Verbal LF:

Degrad [Lat. degradare (to degrade, worsen)]

a. Degrad(clothes) = to wear off

b. Degrad(house) = to become dilapidated

c. Degrad(temper) = to fray

Nominal LF:

Centr [Lat. centrum (the center/culmination of)]

a. Centr(crisis) = the peak (of the crisis)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

b. Centr(desert) = the heart (of the desert)

Furthermore, it is assumed that all languages, in different ways, realize the meanings postulated by LFs and that the main difference lies in the language-specific ways in which the combination of given lexical items is used to arrive at various LF meanings. In this sense, LFs are considered as universal functions capturing the meaning of collocations of words and not only. In this context, they can be used as predictors of words and similar, in intention, with the neural word embeddings algorithms and machine learning approaches as of the frequentists' approaches. In other words, MTT aimed at providing a complete linguistic framework for the mapping from the content or meaning of an utterance to its form or text, with collocations being one particular lexical surface realization. The overall lexicographic goal of MTT has been the creation of so-called Explanatory Combinatorial Dictionaries (ECDs) [8] displaying the combinatorial properties of word combinations in a language.

Another historical motivation for the study of word meaning in terms of collocation and co-occurrence has been provided by clinical phycologists [9]. In their experiments conducted with 1,000 people of varied educational backgrounds and professions, the participants were asked to give the first word that comes to their mind as a result of a stimulus word. The experiments have been repeated and translated in several natural languages and produced interesting human association lists. For instance, the similarity lists, which have been produced for the stimulus words *house* and *home*, respectively, are as follows, in order of descending association strength, from left to right:

- *Home*: {house, family, mother, away, life, parents, help, range, rest, stead}
- *House*: {home, garden, door, boat, chimney, roof, flat, brick, building, bungalow}

A mathematically, however, motivated line of influence on today's computation of relations between words was firstly established by Zelig Harris, who introduced the distributional hypothesis [10]. He stated that *linguistic analysis should be understood in terms of a statistical distribution of components at different hierarchical levels and constructed a practical conception on this topic*. Although Harris believed that language is a system of many levels, in which items at each level are combined according to their local principles of combination, which does not necessarily exclude semantics, was turned towards a more syntactic (formation rules) and logic (transformation rules) interpretation of meaning instead of semantics by focusing on relations between linguistic units. Hence, he hardly escaped the grammatical and lexical collocations as of his predecessors.

It was only a few decades later when these two directions of research (Firth and Harris) converged into an interpretation of meaning in linguistics from a computational point of view. This confluence was made possible by other researchers in the field such as Church, Smadja, et al [11-13]. This new approach was partly derived from psycholinguistic research into word associations and was combined with methods from information theory (mutual information) and computation (co-occurrences). Church applied this to simulate learning on a large corpus of text. They produced simulated knowledge about word associations, which was used to extract lexical and grammatical collocations. He also pointed out other possible applications, especially the solution of polysemy.

In this context, the usage of the term 'word association' indicates a broader meaning. In their examples of automatically computed, strongly associated word pairs, there is a mentioning of semantic relations such as *meronymy*, *hyperonymy* and so forth. Smadja, however, mentions them as examples of where Church's algorithm computed just 'pairs of words' that frequently appear together' [14]. Lin [15] even considers 'doctors' and 'hospitals' as unrelated and thus wrongly computed as significant by Church and Hanks [16], although they stand in a meronymy relation. Nonetheless, other contemporaries, e.g., Dunning [17], improved the mathematical foundation of this research field by introducing the log-likelihood measure. Dunning among the first to coin the term 'statistical text analysis'.

In the era of big data analytics and deep learning, techniques to extract lexical meaning of words from text corpora, questions have risen as to which extent these algorithmic and machine learning approaches are capable of distinguishing between co-occurences and semantic dependencies, which are corpus independent, and those which are corpus dependent. The question also rose as if there is anything else in natural language processing, which goes beyond Deep Learning.

In this paper and in the context of 'statistical text analysis' and deep learning, we will try to give some answers to questions related with the limitations of statistical text analysis and machine learning techniques in regards with the extraction of word associations and computing of semantic similarities. Given also that evaluating the results of semantic similarity algorithms has proven to be quite complicated, as there is no easy way to define a gold standard, we will make an attempt to establish a cross-disciplinary evaluation framework and, therefore, avoid the many different methods of indirect evaluation, which have been used in the past. This framework will be informed by the following approaches: a) linguistics and collocation dictionaries as of the Meaning Text Theory (MTT), b) psychology and human association lists.

The paper is structured as follows: Section 2 provides an overview of the most established algorithmic and machine learning approaches in NLP such as LSA, LDA, Word2vect, GloVe, Deep Learning. These have as common denominators the facts that (a) lexical meaning of words is determined by its surrounding words in a given document or corpus, which, in turn, are defining what is *the context*, (b) words are turned into numbers, in order to enable similarity measurements.

Section 3 provides an evaluation framework by initially discussing some methodologies and principles as derived from past cased studies as an attempt to compare intradisciplinary approaches, e.g., distributional semantics based approaches, as well as some cross-disciplinary ones, e.g., LSA versus human association lists. Subsequently, we embark on our methodology as more holistic approach towards measuring the quality of association lists in that we contrast machine association lists with both MTT based and psychologically induced association lists.

Finally, section 4 discusses the results and draws some first conclusions about the strengths and weaknesses, as well as limitations, of machine association lists. It also attempts to demystify Deep Learning and other contemporary machine learning approaches for NLP paving also the way towards new algorithmic approaches for NL processing and understanding.

2. Overview of algorithmic approaches

2.1 *Computing semantic similarity*

Although it is quite difficult to provide an exhaustive list of related word, we will attempt to discuss the related work alongside three main research directions. As already discussed in the introduction, since the early 1990s, the development of the statistical analysis of natural language has split into three directions. **The first direction** can be viewed as *extraction of collocations*, which was initiated by Church and Smadja [11-13], and continued by Evert and Krenn [18], Seretan [19] and Evert [20]. Main applications of this line of research can be found in translation and language teaching, where it is important to know which expressions are common and which are not possible, in order to avoid typical foreigners' mistakes.

The **second direction** of development can be roughly coined as *extraction of word associations and computation of semantic similarities*. Generally speaking, the main idea has been to (semi-)automatically extract pairs of 'somehow' related or similar words by statistically observing their co-occurrence patterns. The resulting pairs of words of significant co-occurrence, however, are not necessarily idiosyncratic collocations as there are many factors, which can be responsible for the frequent co-occurrence of two words, since word association since this is a rather vague relation allowing for many interpretations.

In this sense, two words might be considered associated with each other in some way. This is also exacerbated by vague definition of context, which may vary from n-gram, i.e., a certain amount of words to the left or right, to the whole document or corpus. Another distinguishing feature has been the way these algorithms group words. This may be a way that is more indicative of syntactic class information, while other algorithms such as Latent Semantic Analysis (LSA) [21] and the topics model, as particularly addressed by the Latent Dirichlet Allocation (LDA) [22], seem to extract structure that might be described as semantic. Still other algorithms such as Hyperspace Analog to Language (HAL) [23] appear to capture a combination of syntactic and semantic information.

 doi:10.20944/preprints201805.0102.v2

The results, however, obtained by algorithms from this field were useful and have therefore been applied in many different applications, such as word sense disambiguation, e.g., [24], word sense discrimination, e.g., [25], or the computation of thesauri, e.g., [26], and to a lesser extent in key word extraction, e.g., [27], text summarization, e.g., [28], and extraction of terminology, e.g., [29].

The **third direction** of development is attributed to the (semi-)automatic extraction of particular linguistic relations (or thesaurus relations), e.g., [30], which are also known as automatic construction of a thesaurus. This line of development has to be distinguished from the other two lines of research in that it introduces a different methodology based on second order statistics, differentiating between syntagmatic and paradigmatic relations [31], context comparisons [32]. Besides, this line of development attempts to give the term 'word association' a more precise definition, which can be used to denote various kinds of linguistic relations, often synonyms, sometimes plain word association (play, soccer) and sometimes other linguistic relations like derivation and hyperonymy, antonyms, qualitative direction of adjectives (negative vs. positive), e.g., [33-34]. Word sense distinction, contrary to word sense disambiguation, e.g., [35], belongs to this area as well, since it describes just another kind of specific relations between words.

In this paper, we will further consider typical approaches and representatives from the **second direction of research**, which is coined as *extraction of word associations and computation of semantic similarities*. This s due to two main reasons: a) most influential and impact creating algorithms can be found in this category, b) strongly related with big data analytics and deep learning. In the following, we will briefly discuss some main representatives of these algorithmic and machine learning approaches in a hope to illustrate the context within which these approaches operate and, consequently, illustrate their limitations.

2.1.1 Memory-based approaches

More specific, memory-based algorithmic approaches take the view that words, which commonly fill similar contexts, are said to have high substitution probabilities and are deemed to be similar [36]. This approach takes the view that sentence processing involves the retrieval of sentence fragments from memory and the alignment of these fragments with the sentence to be interpreted. Retrieval and alignment are achieved using a Bayesian version of String Edit Theory (SET) [37]. In order to employ SET, a matrix of edit operation probabilities is usually induced. Edit operation probabilities can be thought of as the lexical memory of the system, and the substitution probabilities, i.e., the probability that one word can substitute for another, can be thought of as lexical similarities. This procedure, however, involves taking each sentence fragment from a corpus and comparing it against every other sentence fragment. Hence, this procedure is computationally expensive for large corpora where there may be tens of millions of fragments to be compared against each other.

In order to reduce the inherited time complexity, algorithmic approaches appeared, which make a few assumptions and achieve a fast approximation to the generic procedure. The key idea of these algorithms has been to divide the sentence fragments into equivalence classes such that each fragment needs only be compared against those from the same equivalence class rather than the entire corpus [38]. In this context, very high frequency words are used as boundaries of a fragment, which is defined as a sequence of words bounded by these very high frequency words at the beginning and the end of sentence. Subsequently, fragments with the same length and high frequency words form word patterns and belong to the same equivalence class.

For instance, the sentence "THE book showed A picture OF THE author carrying A copy OF THE manuscript." Would be divided into the following fragments:

1. THE book showed A
2. A picture OF THE
3. OF THE author carrying A
4. A copy OF THE
5. OF THE manuscript

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

where the very high frequency words are marked in capital letters. Therefore, the second and fourth fragments would be assigned to the same equivalence class as they contain the same pattern of high frequency words. Consequently, it would be deduced that "picture" and "copy" may substitute for one another. As exemplified by [38], calculating substitution probabilities takes each fragment within an equivalence class and matches it against each other fragment in that class only, not against all possible fragments in a text corpus. The matching strength is the count of the number of words in position that the fragments have in common. This matching strength was then normalized against the total matching strength for all of the fragments within the equivalence class. These retrieval probabilities are then averaged across the instances of each target word appearing in different fragments. For instance, assuming that the following equivalence classes hold

A copy OF THE
A description OF THE
A side OF THE

and

ONTO THE copy
ONTO THE table

The similarity between the words *picture* and *copy* is calculated as being the average retrieval probability of substituting the word *picture* with the *word* copy, i.e., P(<picture, copy>) = (0.5+0.33)/2 = 0.415. This is elaborated on the grounds of the combined matching strength between the fragment “A picture OF THE” and the first equivalent class (e.g., 1 / 3 = 0.33 as of having three high frequency words in common with a class having three other members), as well as between the fragment “ONTO THE picture” and the second equivalence class (e.g., 1 / 2 = 0.5 as of having two common high frequency words in common with a class having two other members).

2.1.1 Distributional semantics

A long tradition in computational linguistics has shown that contextual information provides a good approximation to word meaning, since semantically similar words tend to have similar contextual distributions [39]. In concrete, distributional semantic models (DSMs) use vectors that keep track of the contexts, e.g., co-occurring words, in which target terms appear in a large corpus as proxies for meaning representations, and apply geometric techniques to these vectors to measure the similarity in meaning of the corresponding words.

In this context, vector based approaches take the view that a target word is compared against the vectors for other words in order to determine similarity. For instance, the Pooled Adjacent Context (PAC) model [40] constructs a representation of a word by accumulating frequency counts of the words that appeared in the two positions immediately before and immediately after the target word. The four position vectors created in this way are then concatenated to form the representation of the word. For instance, in the context of the exemplary following windows of text

found a picture of the
found a picture in her
a pretty picture of her

found a copy of a
found a copy below the
destroyed the copy of the

the similarity between *picture* and *copy* would have been calculated by setting two vectors with the frequencies of particular words in two positions left and right of the two words in question. For example, the vector of the word *copy* would be [2 1 0 0 2 1 2 0 1 2 0 1] for all words appearing at positions -1, -2, 1, 2 in all these text windows.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

Latent Semantic Analysis (LSA)

LSA [21] takes the idea of extracting lexical meaning of words from the sentential context a little bit further. The underlying idea is that the aggregate of all the word contexts, in which a given word does and does not appear, provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. It has been claimed that LSA reflects on human knowledge, which may have been established in a variety of ways. Analytical studies in the past showed that LSA scores overlap those of humans on standard vocabulary and subject matter tests. LSA is also known to mimic human word sorting and category judgments, as well as the way it simulates word–word and passage–word lexical priming data. Finally, it has been reported that it accurately estimates passage coherence, learnability of passages by individual students, and the quality and quantity of knowledge contained in an essay.

LSA relies on the follows method. After processing a large sample of machine-readable language, LSA represents the words used in it, and any set of these words, such as a sentence, paragraph, or essay, as points in a very high (e.g. 50-1,500) dimensional “semantic space”. LSA is closely related to neural net models, but is based on singular value decomposition (SVD), a mathematical matrix decomposition technique closely akin to factor analysis that is applicable to text corpora approaching the volume of relevant language experienced by people.

More specific, in SVD a rectangular matrix is decomposed into the product of three other matrices. One component matrix describes the original row entities as vectors of derived orthogonal factor values, another describes the original column entities in the same way, and the third is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed. There is a mathematical proof that any matrix can be so decomposed perfectly, using no more factors than the smallest dimension of the original matrix.

It is worth noting that similarity estimates derived by LSA are not simple contiguity frequencies, co-occurrence counts, or correlations in usage, as of the previous approaches, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations, e.g., the phrase “Latent Semantic”. As a consequence, these estimates are often much better predictors of human meaning-based judgments and performance than are the surface level contingencies, some of which have been rejected by linguists as the basis of language phenomena.

LSA, however, induces its representations of the meaning of words and passages from analysis of text alone. None of its knowledge comes directly from perceptual information about the physical world, from instinct, or from experiential intercourse with bodily functions, feelings and intentions. Thus while LSA’s potential knowledge is surely imperfect, it is believed that it can offer a close enough approximation to people’s knowledge to underwrite theories and tests of theories of cognition.

Nonetheless, LSA has some additional limitations. It makes no use of word order, thus of syntactic relations or logic, or of morphology. LSA also differs from some statistical approaches in two significant respects. Firstly, the input data "associations" from which LSA induces representations are between unitary expressions of meaning, i.e., words and complete meaningful utterances in which they occur rather than between successive words. LSA uses as its initial data not just the summed contiguous pairwise (or tuple-wise) co-occurrences of words but the detailed patterns of occurrences of very many words over very large numbers of local meaning-bearing contexts, such as sentences or paragraphs, treated as unitary wholes. Thus it skips over how the order of words produces the meaning of a sentence to capture only how differences in word choice and differences in passage meanings are related.

Another way to think of this is that LSA represents the meaning of a word as a kind of average of the meaning of all the passages in which it appears, and the meaning of a passage as a kind of average of the meaning of all the words it contains.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

2.1.2 Latent Dirichlet Allocation

A topic model is a kind of a probabilistic generative model that has been used widely in the field of computer science with a specific focus on text mining and information retrieval in recent years. Since this model was first proposed, it has received a lot of attention and gained widespread interest among researchers in many research fields. The origin of a topic model is latent semantic indexing (LSI) [41]; it has served as the basis for the development of a topic model. Nevertheless, LSI is not a probabilistic model; therefore, it is not an authentic topic model. Based on LSI, probabilistic latent semantic analysis (PLSA) [42] was proposed by Hofmann and is a genuine topic model. Published after PLSA, Latent Dirichlet Allocation (LDA) [22] is treating sentential context in a rather different way than LSA in that it focusses more on associating a document with a topic such as *cute animals*.

Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" may appear more often in documents about *cure animals*. Moreover, a topic model can be represented as a graphical model, or probabilistic graphical model (PGM), or structured probabilistic model. In that sense, a graph expresses the conditional dependence structure between random variables.

More formally, LDA is conceived as a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. LDA often relies on efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation [22].

In order to exemplify LDA, let us assume that we have the following set of sentences:

- I like to eat broccoli and bananas.
- I ate a banana and spinach smoothie for breakfast.
- Chinchillas and kittens are cute.
- My sister adopted a kitten yesterday.
- Look at this cute hamster munching on a piece of broccoli.

LDA may have allocated the following probabilities:

Sentences 1 and 2: 100% Topic A (food)
Sentences 3 and 4: 100% Topic B (cute animals)
Sentence 5: 60% Topic A, 40% Topic B
Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching
Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster

In that sense, a document D, which may contain these sentences will be represented with conditional probabilities allocated to topics A and B. In other words, assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

From a machine learning point of view, one has to choose some fixed number of K topics to discover for a given set of documents as you want to use LDA to learn the topic representation of each document and the words associated to each topic. Generally speaking, the algorithm(s) go through each document and randomly assign each word in the document to one of the K topics. Consequently, in order to improve these assignments, for each word w in a document d, and for each topic t, LDA computes two things: 1) p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and 2) p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w. Subsequently, a new topic is reassigned to w, where the topic t is chosen with probability p(topic t | document d) * p(word w | topic t). Repeating the previous step a large number of times, the algorithm eventually reaches a roughly steady state where the assignments are pretty good.

The main disadvantages being reported are associated with the question "how hard it is to know when LDA is working", since topics are soft clusters so there is no objective metric to say "this is the best choice" of hyperparameters. Metrics like perplexity (how well the model explains the data) can be applied if the learning is working. They are, however, poor indicators of the overall quality of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

model. For example, you could have a model with very low perplexity, but whose topics are not very informative. Furthermore, LDA and most of its variants rely on a Bag of Words (BoW) approach. In a sense, it still treats documents as a bag of words and the exchangeability of words and documents could be called the basic assumptions of a topic model. These assumptions are available in both PLSA and LDA. Nevertheless, in several variants of topic models, a basic assumption was relaxed.

In this context, topic modeling with LDA and its variants does not address the lexical meaning of words as such. It is more seen as a side effect. Moreover, it became obvious that relaxing the basic assumption of LDA or PLSA is a desirable approach, since the availability of many other a priori pieces of information, such as documents' interactions, the order of words, and knowledge on the biology domain, play an important role as well. In addition, there is significant motivation to reduce the time taken to learn topic models for very large data, for instance, in biological data.

2.2. Artificial Neural Networks (ANNs)

As already discussed in [44], ANNs are robust learning models that are about precisely assigning weights across many levels. They are broadly divided into two types of ANN architectures: those that can be feed-forward networks and those Recurrent (or Recursive) Neural Networks (RNNs) [45]. Feed-forward architecture consists of fully connected network layers. The RNNs model, on the other hand, consist of a fully linked circle of neurons connected for the purpose of back-propagation algorithm implementation. ANNs applied to NLP tasks consider syntax features as part of semantic analysis [46]. New neural network learning models have been proposed that can be applied to different natural language tasks, such as semantic role labelling and Named Entity Recognition [47]. The advantage of these approaches is to avoid the need for prior knowledge and task specific engineering interventions. ANN models have achieved an efficient performance in tagging systems with low computational requirements [48].

Word2vec

Word2vec [49] can be viewed as a two-layer neural network that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. Google calls it "an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words."

While Word2vec is not a deep neural network (*see next subsection for more details about deep learning architectures*), it turns text into a numerical form that deep networks can understand. In that sense, Word2Vec is a particularly computationally efficient predictive model for learning word embeddings from raw text. For instance, given the sentence "The cat was sitting on the …", Word2vec is likely to predict the next word being "mat". Therefore, highly accurate guesses about a word's meaning can be made, which are based on past appearances. Those guesses can be used to establish a word's association with other words (e.g. "man" is to "boy" what "woman" is to "girl"), or cluster documents and classify them by topic.

The output of the Word2vec neural network is a vocabulary in which each item has a vector attached to it, which can be fed into a deep-learning network or simply queried to detect relationships between words. For instance, a list of words associated with "Sweden" using Word2vec, in order of proximity, is given as of the following vector:

Word	Cosine distance
norway	0.760124
denmark	0.715460
finland	0.620022
switzerland	0.588132
belgium	0.585835
netherlands	0.574631
iceland	0.562368
estonia	0.547621
slovenia	0.531408

The similarity of the word “Sweden” to other words is measured as the cosine similarity between word vectors. Zero similarity is expressed as a 90 degree angle, while total similarity of 1 is a 0 degree angle. For instance, a complete overlap; i.e., Sweden equals Sweden, gives a total similarity of 1, while *Norway* has a cosine distance of 0.760124 from Sweden, the highest of any other country.

The vectors being used to represent words are called *neural word embeddings*, and representations are strange; one thing describes another, even though those two things are radically different. Word2vec comes in two flavours, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. Algorithmically, these models are similar, except that CBOW predicts target words (e.g. 'mat') from source context words ('the cat sits on the'), while the skip-gram does the inverse and predicts source context-words from the target words. This inversion might seem like an arbitrary choice, but statistically it has the effect that CBOW smooths over a lot of the distributional information (by treating an entire context as one observation). For the most part, this turns out to be a useful thing for smaller datasets. However, skip-gram treats each context-target pair as a new observation, and this tends to do better when we have larger datasets.

In a nutshell, similar things and ideas are shown to be “close” in that their relative meanings have been translated to measurable distances. Similarity is the basis of many associations that Word2vec can learn. Since words are represented as vectors, powerful mathematical operations can be applied. It was recently shown that the word vectors capture many linguistic regularities, for example vector operations such as *vector('Paris') - vector('France') + vector('Italy')* results in a vector that is very close to *vector('Rome'), and vector('king') - vector('man') + vector('woman')* is close to vector('queen'). Despite these information retrieval operations, Word2vec is predominantly a "context predictive" model, which earn their vectors in order to improve the loss of predicting the target words from the context words given the vector representations.

Global Vectors (GloVe)

Similar to Word2vec approach, GloVe [50] is another unsupervised learning algorithm for obtaining vector representations for words. The main difference, however, is that training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space. In that sense, GloVe is usually classified as *count-based model,* which learn the vectors by essentially doing dimensionality reduction on the co-occurrence counts matrix. Firstly, a large matrix of words x in context y is constructed based on co-occurrence information, i.e., for each "word" (the rows), the learning algorithm counts how frequently we see this word in some "context" (the columns) in a large corpus. The number of "contexts" is, of course, large, since it is essentially combinatorial. Hence, factorization of the matrix is applied in order to yield a lower-dimensional matrix, where each row now yields a vector representation for each word.

 doi:10.20944/preprints201805.0102.v2

Deep Learning Architectures

Deep learning is essentially a bigger take on the neural network models that have been around for some time. It is attribute to Geoffrey Hinton and his first attempts to develop an image classification algorithm. It is, however, particularly useful for analyzing, audio, text, genomic and other multidimensional data that does not lend itself well to traditional machine learning techniques.

Word vectors to be used for similarity measures, as previously discussed, can be learned by applying Deep Learning (DL) based architectures as well. DL, as a yet another ANN based architecture, involves multiple data processing layers, which allow the machine to learn from data through various levels of abstraction for a specific task without human interference or previously captured knowledge. Therefore, one could classify DL as unsupervised Machine Learning (ML) approach. Investigating the suitability of DL approaches for NLP tasks has gained much attention from the ML and NLP research communities, as they have achieved good results in solving bottleneck problems [51].

These techniques have had great success in different NLP tasks, from low level (character level) to high level (sentence level) analysis, for instance, sentence modelling [52], Semantic Role Labelling [48], Named Entity Recognition [53], Question Answering [54], text categorization [55], opinion expression [56], and Machine Translation [57].

More specific, since Deep Learning is based on Convolutional Neural Network (CNN) architectures, which has been around for more than three decades, CNNs have been applied as a non-linear function over a sequence of words, by sliding a window over the sentences. This has been the key advantage of using CNNs architecture for NLP tasks. This function, which is also called a 'filter', mutates the input (k-word window) into a d-dimensional vector that consists of the significant characteristic of the words in the window. Then, a pooling operation is applied to integrate the vectors, resulting from the different channels, into a single n-dimensional vector. This is done by considering the maximum value or the average value for each level across the different windows to capture the important features, or at least the positions of these features. For example, **Error! Reference source not found.** gives an illustration of the CNNs' structure where each filter executes convolution on the input, in this case a sentence matrix, and then produces feature maps, hence it is showing two possible outputs. This example is used in the sentence classification model.

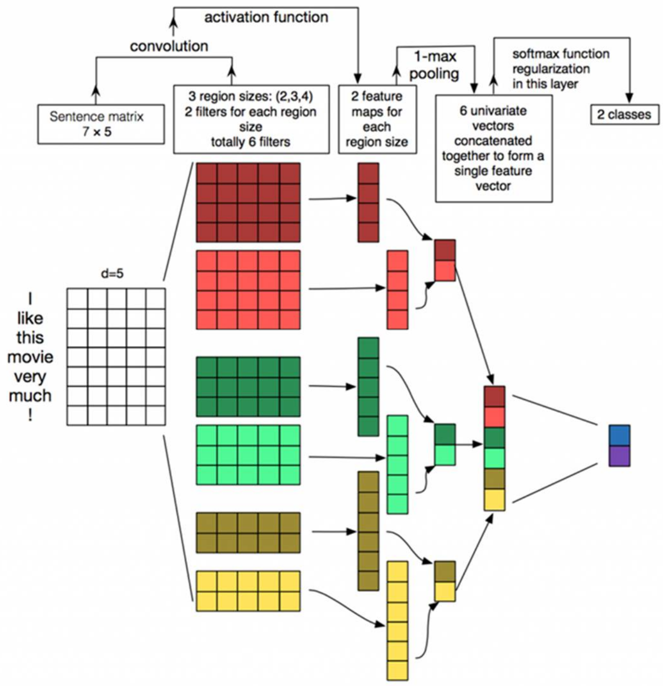

Figure 1: Model of three filter division sizes (2, 3 and 4) of CNNs architecture for sentence classification. (Source: [61])

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

A new convolutional latent semantic approach for vector representation learning [58] uses CNNs to deal with ambiguity problems in semantic clustering for short text. However, this model can work appropriately for long text as well [59]. CNNs are proposed for sentiment analysis of short texts that learn features of the text from low levels (characters) to high levels (sentences) to classify sentences in positive or negative prediction analysis. However, this approach can be used for different sentence sizes [60].

In a nutshell, building a machine-learning system with features extraction requires specific domain expertise in order to design a classifier model for transforming the raw data into internal representation inputs or vectors. These methods are called representation learning (RL) in which the model automatically feeds in raw data to detect the needed representation. In particular, the ability to precisely represent words, phrases, sentences (statement or question) or paragraphs, and the relational classifications between them, is essential to language understanding.

3. Evaluation methodology

Evaluating the results of semantic similarity algorithms for the extraction of word associations has proven to be quite complicated. There is mainly due to the following reasons:

- There is no easy way to define a gold standard, and therefore many different methods of indirect evaluation have been used.
- The notion of 'context' is scattered across a broad spectrum ranging from n-gram models, where context is simply an n-gram, to windowing models, where context is defined as number of words to the left and to the right of the observed word, to a notion of context which means the whole text in which the observed word occurs.
- The type of the word association being targeted. Roughly speaking, three types of associations may be targeted: *syntactic structure, semantic structure, associative structure*. The latter is captured in two main flavors:
    - *syntagmatic associations* (e.g., run-fast), which are thought to be acquired as consequence of words appearing in succession in the experience of the subject;
    - paradigmatic associations (e.g., run-walk), which are thought to occur as consequence of experiencing words in similar sentential contexts.

Further humbling aspects for easing off the evaluation complexity of these algorithmic approaches have been the variety of algorithms (e.g., type 0, type 1, type 2, type 4), as well as the ways the strength of an association is being measured (e.g., from mutual information, to comparisons of binary and real-valued vectors).

Despite the inherited complexity of these evaluation methods, systematic comparisons of algorithms and models have been attempted in the past. For instance, [62] have attempted to quantitatively contrast the abilities of these algorithms to capture all three types of associations, namely, syntactic, semantic and associative information. Much, however, remains to be done to characterize the type of word association each of these algorithms acquire. Moreover, [63] carried out a systematic comparison between context-predicting and context-counting semantic vector approaches, which underpins the differentiation between Word2vec and GloVe semantic vectors. This evaluation, however, does not target all three types of associations and does not give a clear definition of the term 'word association'.

The most promising and most comparable evaluation is one using large manually crafted knowledge sources such as Roget's Thesaurus [64], WordNet [65-66] or GermaNet for German [67] as a gold standard. Unfortunately, again, evaluations using these sources can be done in many different ways, crippling comparability. A standardized tool set or instance is needed.

3.1. Our methodological approach

After considering the various evaluation methods and the inherited complexity of evaluating the quality of extracted word relations, a conclusion was drawn that for the purposes of this study: the gold standard should probably be

- either a collocations dictionary like BBI Combinatory Dictionary of English and Explanatory Combinatorial Dictionaries (ECDs),
- or a semantic net like WordNet.

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. Apart from gold standards, however, the following pillars expanded our evaluation methodology: *psycholinguistic association or priming experiments, vocabulary tests, application-based evaluations, evaluation by using artificial synonyms.*

Association or priming paradigms [68] can be used to evaluate the results of the algorithms by comparing them with data obtained from human subjects in psycholinguistic experiments. Suitable are association or priming experiments, where subjects are asked to name rapidly some semantically close words after being presented with the stimulus word. The list of most frequently named words can then be compared with the lists obtained automatically.

A vocabulary test usually comprises a question and a multiple-choice answer. If both are electronically available, the test can be used quite straightforwardly to evaluate word similarity computation methods. TOEFL, i.e., Test of English as a Foreign Language, has been used as one the tests comprising 80 test items. This kind of evaluation has been used by many authors, such as [69], [21], [70-71].

Application-based evaluation is the indirect method of evaluating results of a knowledge extraction algorithm by putting the extracted knowledge into use and observing how well the application using this knowledge performs. One of the most interesting approaches, however, is the use of artificial items. The main idea for testing synonymy is to choose randomly one part of occurrences of a word and replace the word by a pseudo-word while keeping the other part. It is then possible to measure how often the pseudo-words are extracted as synonyms of the words that have been retained.

4. Preliminary results and discussion

Our comparison study is based on some preliminary results, which have been the outcome of the application of *Deep Learning* techniques in order to improve the extracted Word2vec model as a means to compute vector representations of words. For the sake of this comparison study, we will refer to the Eclipse *Deeplearning4j* as an open-source, distributed deep-learning project in Java and Scala spearheaded by the people at *Skymind*, a San Francisco-based business intelligence and enterprise software firm. *Deeplearning4j* implements a distributed form of Word2vec for Java and Scala, which works on Spark with GPUs. The extracted word associations, as listed in Table *1*, which rely on the trained Word2vec model, have been trained on the Google News vocabulary, which you can import and play with from the Google News Corpus Model (GoogleNews-vectors-negative300.bin.gz, 1,5 GB).

For the interpretation of the word associations, the following notations hold: where : means "is to" and : : means "as". For instance, "Rome is to Italy as Beijing is to China" = Rome:Italy::Beijing:China

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

Table 1: Arrays of extracted word associations

1	king:queen::man:[woman, Attempted abduction, teenager, girl]
2	China:Taiwan::Russia:[Ukraine, Moscow, Moldova, Armenia]
3	house:roof::castle:[dome, bell_tower, spire, crenellations, turrets]
4	knee:leg::elbow:[forearm, arm, ulna_bone]
5	New York Times:Sulzberger::Fox:[Murdoch, Chernin, Bancroft, Ailes]
6	love:indifference::fear:[apathy, callousness, timidity, helplessness, inaction]
7	Donald Trump:Republican::Barack Obama:[Democratic, GOP, Democrats, McCain]
8	monkey:human::dinosaur:[fossil, fossilized, Ice_Age_mammals, fossilization]
9	building:architect::software:[programmer, SecurityCenter, WinPcap]

Noteworthy is that the Word2vec algorithm has never been taught a single rule of English syntax. It knows nothing about the world, and is unassociated with any rules-based symbolic logic or knowledge graph.

Despite the limited number of extracted word associations, these results seem to confirm that the extracted associations do not capture all three types of associations, namely, *syntactic, semantic and associative information*. and does not give a clear definition of the term 'word association'. For instance, the word associations *King* - *Queen* and *Man* – *Woman* do not provide any clue about the type of association holding between these words. There is, however, a *semantic structure* as a type of association being derived implicitly from the relationship "as" or "same as" holding between the pairs of words {King, Queen} and {Man, Woman}: a *King is a Man*, a *Queen is a Woman*. Even so, there is no reference to whether this semantic structure is a *hyperonymy*, a semantic relation between a more general word and a more specific word, or *meronymy*, a semantic relation, which refers to a part of a whole and usually characterized as "part-of" relationship.

Moreover, there is no such a thing as a pattern of semantic relationships emerging from the first pairs of word associations at both sides of the notation : :. For instance, neither a *hyperonymy* nor a *meronymy* seem to be the case for the other word associations on the list, e.g., {monkey, human} and {dinosaur, fossil}, as one cannot infer any relationship between *monkey* and *dinosaur*, or between *human* and *fossil*. Even if we succeed to identify a pattern of relations, i.e., *two large countries and their small, estranged neighbors*, such as those emerging from the second row word associations on the list, we cannot emerge victorious with a pattern of semantic relations when we do the same with the eighth row word associations. We will stumble upon questions as to which extent *humans should be considered as fossilized monkeys*, or *humans are what's left over from monkeys*, or *humans are the species that beat monkeys* just as *Ice Age mammals beat dinosaurs*.

An interesting observation has also been as to which extent a holding relationship between two words could imply the same relationship or association type on the other side of the notation : :. For instance, as of the ninth row word associations, and assuming that an *architect is-the-designer of a building*, can we imply that a *programmer is-the-designer of a software*? At first glance, it looks like that such a pattern does hold as in most of the cases a well predicted relationship seem to be holding on the other side of the notation : :. There is, however, a notorious difficulty in identifying what are exactly these relations, which can hold on both sides, hence, inferring the one will imply the other.

Moreover, [63] carried out a systematic comparison between context-predicting and context-counting semantic vector approaches, which underpins the differentiation between Word2vec and GloVe semantic vectors.

4.1 Comparisons with a golden standard (lexicography)

As indicated in section 3.1, we used as a golden standard the English Collocations Dictionary which is available online at the URL www.ozdic.com, as well as the online version of WordNet 3.1 available online at the URL https://wordnet.princeton.edu/ The intention has been to confirm whether the extracted word associations, for all pairs of words, can be replicated by the collocations

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

dictionaries, as well as whether the same semantic relationship, be it semantic or lexical, holds across both sides of the notation : : In the following, the results of these comparisons are presented for each list of extracted word associations. All potential relations have been checked bi-directionally, e.g., entries have been both words *King* and *Queen*.

Having checked all word entries, we identified two lists, 5 and 7, which have no single collocation. Both lists do predominantly refer to named entities, e.g., *Donald Trump, New York Times*. Besides, From the total of thirty (30) pairs of associated words, we could identify seventeen (17) collocations in the dictionary, i.e., slightly over 50% of all possible word associations. The following Table 2 summarises the identified collocations together with the potential relations holding between them.

Table 2: Identified collocations for the English language as of WordNet and ozdic.com

Extracted word associations	**Source: www.ozdic.com**	**Source: WordNet 3.1**
King - Queen	Wife of	Wife or widow of
Man – Woman	-	Wife / Mistress / Girlfriend
Russia – Ukraine	-	Former parts of USSR
Russia – Moscow	-	Part of / capital of
China - Taiwan	-	Part of / governed by
House – roof	Under your	-
Castle – bell tower	Castle + noun / flanked	
Castle - turrets	Adjective + Castle	
Castle – Crenellation	-	Part of (meronymy)
Knee - leg	Below the / amputated below the	Part of (meronymy)
Elbow - arm	Below the /	Part of (meronymy)
Elbow – forearm	-	Part of (meronymy)
Elbow – ulna bone	-	Elbow bone as a synonym to ulna bone
Love - indifference	-	Causing (love -> indifference)
Monkey - Human	-	Both being part of experiments
Building - Architect	-	Engaged in / building
Software - programmer	-	Builds / designs / writes / tests

Subsequently, we tried to answer the question whether the indicative relations, as indicated by both online resources for the lexical and semantic word meaning, can be projected on the other side of the notation : :. It turned out that almost all of the above relations can be imposed on one or more word associations on either side of the notation : :. For instance, it is perfectly acceptable to impose the relation "wife of" on the word associations {man, woman} and {man, girl}, as well as the relations "amputated below the" or "being part of" for both pairs {knee, leg} and {elbow, arm}. The same holds for the pairs of words {house, roof} and {castle, crenellations}, in terms of the relation "part of", as well as for the pairs of words {house, roof} and {castle, turrets}, since the expression "roofed house" and "turreted castle" are both meaningful. In some cases, however, e.g., {monkey, human}, the indicative relation cannot be imposed on the other part of the notation : :.

Overall, it seems to be indicative that, despite the notorious difficulty to extract the type of association or the relation holding between the pairs of words, some of these word associations do, indeed, make sense according with the lexicographic and semantic meaning of words as indicated by the two lexicographic resources. Furthermore, in some cases, the underpinning relation is rather

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

vague and uncertain as the case with sentiments, e.g., in the array *fear:[apathy, callousness, timidity, helplessness, inaction]*.

On the other hand, considering the arrays

Donald Trump:Republican::Barack Obama:[Democratic, GOP, Democrats, McCain]

monkey:human::dinosaur:[fossil, fossilized, Ice_Age_mammals, fossilization]

there may be some interesting relations, which remain hidden. For instance, given the fact that Obama and McCain were rivals, it may be interesting to investigate whether the relation "rivalry" may also hold between *Donald Trump* and the ideal *Republican*. In addition, the one plausible relation between *humans* and *monkeys* may be *that humans is the species that beat monkeys* just as *ice age mammals beat dinosaurs*.

4.2 Comparisons with results from psycholinguistic experiments

Although it is notoriously difficult to get access to results from psycholinguistic experiments, for the sake of our comparison study, we will mainly refer to results published in [9, 72] and the *Kent-Rosanoff Word Association Test* in order to study word association norms as a function of age. The experiment has been conducted with 738 subjects from 18 to 87 years of age from various occupations and from various parts of the country. The experiment was meant to study the strength of a word association as a function of age, in terms of a stimulus and response words. For instance, "drinking" as a response to the stimulus word "eating". Consequently, percentages of subjects responding to 100 common word associates for three age groups: Group A: (ages 18-33 years, N= 373), Group B (ages 34-49, N = 205) and Group C (ages 50-87, N = 160).

Despite the idiosyncratic nature of this experiment and in order to avoid drawing false conclusions, we restricted ourselves in checking for common word entries in the list of 99 words as of [72]. Our comparisons verified that it is difficult to infer any semantic or lexical relations holding among the associated words. Hence, from this comparison, there is no directly added value in predicting what the potential relation may be, or whether the "same as" predicate on both sides of the notation : : can be added.

It has been revealed, however, that few of the word associations in our nine (9) arrays of Table *1* do also exist in the results of this experiment. For instance, the associations between *man* and *woman*, *kind* and *queen*, could also be confirmed. The most revealing aspect, however, has been that associations within the same array of associated words, such as between *woman* and *girl* could be unveiled by the entries in the list of 99 words [72]. This may, in turn, indicate, the associations may be transitive as well. For instance, the association between *man* and *girl* may be the result of the associations between *man* and *woman*, as well as *woman* and *girl*.

4. General discussion

In this paper, we discuss some preliminary results and emerging trends and how they can be interpreted in perspective of previous studies, including our own comparisons. The main working hypothesis has been the question(s) as to what are the limitations of *Deep Learning (DL)*, not only for the extraction of word meaning in natural language processing, but also for the extraction of meaningful associations among objects or entities, in general.

The experimental design addressed primarily a DL framework for the following main reasons: a) to demystify the prowess of this ANN based architecture in its capacity to computationally recognize and understand in terms of interpreting associations between words, b) to act as a typical, up to date, representative of machine learning algorithms for natural language processing and understanding, c) to unveil future research directions, d) to establish an evaluation framework for future reference.

Therefore, it is this broader context within which our findings and comparison results should be interpreted, although rather limited than with some statistical significance. Nevertheless, the following major patterns, and implicitly future research directions, could be unleashed:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

- The notorious difficulty of DL, in particular, and all statistics, vector space based algorithms, in general, to infer the type of association or the exact relation underpinning a word association. In other words, this seems to be still an open research question for all frequentists' approaches relying on turning words into numbers, in order to make them comparable.
- This also applies to *Latent Semantic Analysis (LSA)* as reportedly being very close to human judgements about word associations. However, this is very similar with comparisons made against results from psycholinguistic experiments, which may confirm the strength of a word association, but not extract the type of the association or relation being implied.
- Despite this inadequacy, it can also be confirmed the surprising superiority of these approaches to extract strong word associations, even if the underpinning relation is an unknown variable. In other words, what is being extracted seem to be strongly related, however, without knowing how.

As far as the evaluation methodology is concerned, the following key problems, or context, could be confirmed:

- There is no easy way to define a gold standard, and therefore many different methods of indirect evaluation have been used. In our case, we used as gold standard two resources: the semantic net WordNet and the collocations dictionary for the English language. As of our results, it became apparent that identifying the same collocation in both resources is rarely the case. WordNet, however, seems to provide a more comprehensive and complete structure of lexical and semantic relations for English words.
- In any case and in order to cope with the inherited heterogeneity of these resources, we restricted ourselves in identifying *any collocation*, i.e., mentioning both words in the same lexicographical context, as well as to simplify deriving a potential relation.
- The notion of 'context' also emerged in that the findings and comparison results are attributed to word associations extracted from an, admittedly, large corpus of Google News. Despite that one may argue the findings and comparison results do refer to this specific domain, there are two main lines of thought emerging as well: the doubt that learning and training vector space models with other domains of discourse will extract the type of association or relation holding between words, since these are all turned, more or less, in frequencies and numbers.
- In order to avoid the dilemma of which association type, *syntactic structure, semantic structure, associative structure,* should be targeted, we took a more generic approach in that any collocation would matter.
- Finally, ideally speaking, we should evaluate the findings, i.e., extracted word association and meaning, by taking a more holistic approach. In other words, we should also consider, in addition to the chosen gold standards as the result of lexicographers and psycholinguistic experiments, admittedly, of limited scope, word associations as derived from more experiments such as *vocabulary tests, e.g., TOEFL, application-based evaluations, evaluation by using artificial synonyms.*

As far as these evaluation resources are concerned, the following problems and limitations could also be confirmed:

- Psycholinguistic experiments as such are very costly, especially, if they should be applied to large evaluations instead of small samples as done usually. Therefore, it is very probable that the evaluation results may not be representative. Besides, it may not be easily possible for other researchers to reproduce these experiments and validate the results.
- Using vocabulary tests sounds an interesting option, however, testing against only 80 items poses the problem of whether the results will be representative. In such a case overtraining (by fitting thresholds) can occur very fast. Besides, these tests target only synonymy. Hence, these tests can indicate how good the word associations may be, however, not what is exactly the nature of the underpinning linguistic relation or association type.

- Application-based evaluation, as an indirect method of evaluating results of a knowledge extraction algorithm, sounds like another viable evaluation option, since this puts the extracted knowledge into use and observes how well the application using this knowledge performs. In this context, the reviewed algorithmic approaches for corpus based, word meaning extraction, may be positively evaluated in their use by contemporary search engines and information retrieval tasks, however, negatively in the context of knowledge engineering and, particularly, in the context of extracting a knowledge graph or ontology. This is due to the fact that in the context of information retrieval and Web search, the type of relation easily implied is *synonymy*.
- One of the most interesting approaches to evaluating automatic extraction algorithms is by using artificial items. The idea for testing *synonymy* is to choose randomly one part of occurrences of a word and replace the word by a pseudo-word while keeping the other part. Hence, perfectly artificial synonyms are created. It is then possible to measure how often the pseudo-words are extracted as synonyms of the words that have been retained. Due to the difficulty we faced with the creation of artificial antonyms, meronyms or other linguistically related words, and the entrapment imposed by inflicted biases, this evaluation has been left as future work.

5. Conclusions

This paper has been incentivized by the question what do we really learn when we apply state of the art machine learning and statistics based algorithms towards extraction of word associations and, implicitly, contextual word meaning from text corpora. Although the experimental results are preliminary and the comparisons, perhaps, of limited scope, the contribution to knowledge may be sought after in some of the following aspects: a) *confirming the lack of extracted types of association, be them structural, semantic or associative, or specific relations holding among words, despite the fact that state-of-the-art machine learning techniques seem to be strengthening the nature of a word association,* b) *the inherited complexity of an evaluation framework for this purposes due to many reasons ranging from the definition of equivalent contexts to categorizing of algorithms in terms of what type of association is concerned, to lack or difficulty of access to word association lists produced by other human centered efforts and experiments.* Nonetheless, we put the emphasis on open access data and reproducible results by addressing publicly available software and data.

In the future, we will keep on expanding our experiments, not only in terms of producing more data and comparisons, but also in terms of designing and implementing machine learning architectures, which are more keen on extraction of meaningful associations or relations underpinning an extracted word association. This approach will be informed by recent advances and lessons learned in cognitive sciences and human-like robot learning [73], where a robot learns elements of its semantic and episodic memory through language interaction with people. This human-like learning can happen when we extract, represent and reason over the meaning of the user's natural language utterances.

References

1. Firth, J. R. *Papers in Linguistics 1934 – 1951*; Oxford University Press: London, U.K., 1957
2. Manning, C. D.; Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*; MIT Press.
3. Benson, M.; Benson, E.; Ilson, R. *The BBI combinatory dictionary of English: A guide to word combinations*; John Benjamins: Amsterdam, 1986
4. Benson, M. The structure of the collocational dictionary. *International Journal of Lexicography* 1989, 2(1), 1–14.
5. Benson, M. Collocations and general-purpose dictionaries. *International Journal of Lexicography* 1990, 3(1), 23–35.
6. Mel'˘cuk, I. Lexical functions: A tool for the description of lexical relations in a lexicon. In *Lexical Functions in Lexicography and Natural Language Processing*; Wanner, L., Eds.; John Benjamins: Amsterdam, 1996; pp. 23–54.

7. Mel’ˇcuk, I. Collocations and lexical functions. In *Phraseology: Theory, Analysis, and Applications*; Cowie, A., Ed., Clarendon Press: Oxford, 1998; pp. 23–54.
8. Bartsch, S. Structural and Functional Properties of Collocations in English – a corpus study on lexical and pragmatic constraints on lexical co-occurrence; Gunter Narr Verlag: Tübingen, 2004
9. Kent, G.; Rosanoff, A.J. A study of association in insanity. *American Journal of Insanity* 1910, 67, 317-390
10. Zeelig, H. S. *Mathematical Structures of Language*; Wiley: New York, 1968
11. Choueka, Y.; Klein, S. T.; Neuwitz, E. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing* 1983, 34–38.
12. Church, K. W.; Gale, W. A., Hanks, P.; Hindle, D. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting On-Line Resources to Build up a Lexicon*; Uri Zernik, Ed.; Lawrence Erlbaum: Hillsdale, NJ., 1991; pp. 115–164
13. Smadja, F. Macro-coding the lexicon with co-occurrence knowledge. In *Proceedings of the First International Lexical Acquisition Workshop*; Zernik, U., Ed.; 1989
14. Smadja, F. A. Retrieving collocations from text: Xtract. *Computational Linguistics* 1993, 19(1), 143–177.
15. Lin, D. Extracting collocations from text corpora. In *CompuTerm ’98 – Proceedings of the 1st Workshop on Computational Terminology*; Montreal, Quebec, Canada, 1998; pp. 57–63.
16. Church, K. W.; Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics* 1990, 16(1), 22–29.
17. Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 1993, 19(1), 61–74.
18. Evert, S.; Krenn, B. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, 2001, Toulouse, France, pp. 188–195
19. Seretan, M.-V. Syntactic and Semantic Oriented Corpus Investigation for Collocation Extraction, Translation and Generation. Ph.D. thesis, Language Technology Laboratory, Department of Linguistics, Faculty of Arts, University of Geneva, 2003.
20. Evert, S. The Statistics of Word Cooccurrences: Word Pairs and Collocations, Ph.D. thesis, University of Stuttgart, 2005
21. Landauer, T. K.; Dumais, S. T. A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 1997, 104(2), 211–240.
22. Blei, D.M.; Ng, A. Y.; Jordan, M. I.; Latent Dirichlet Allocation, *Journal of Machine Learning Research* 2003, 3, 993-1022
23. Lund, K.; Burgess, C. Producing high dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers* 1996, 28, 203-208.
24. Pantel, P.; Lin, D. 2000. Word-for-word glossing with contextually similar words. In *Proc. of the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics*; 2000, Seattle, USA, pp. 78–85
25. Schütze, H. Automatic word sense discrimination. *Computational Linguistics* 1998, 24, 97–124.
26. Grefenstette, G. *Explorations in Automatic Thesaurus Discovery;* Kluwer Academic Press: Boston, 1994
27. Matsumura, N.; Ohsawa, Y.; Ishizuka, M. PAI: automatic indexing for extracting asserted keywords from a document. *New Generation Computing* 2003, 21(1), 37–47
28. Salton, G.; Singhal, A.; Mitra, M.; Buckley, C. Automatic text structuring and summarization. *Information Processing and Management* 1997, 33(2), 193–207.
29. Witschel, F. Terminology extraction and automatic indexing - comparison and qualitative evaluation of methods. In *Proc. of Terminology and Knowledge Engineering;* 2005
30. Ruge, G. Automatic detection of thesaurus relations for information retrieval applications. In *Foundations of Computer Science: Potential - Theory – Cognition*; Freksa, C., Jantzen, M., Valk, R., Eds.; Springer-Verlag: Heidelberg; pp. 499–506.
31. Rapp, R. The computation of word associations. In *Proceedings of COLING-02*; 2002, Taipei, Taiwan.
32. Biemann, C.; Bordag, S.; Heyer, G.; Quasthoff, Wolff, C. Language-independent methods for compiling monolingual lexical data. In *Proceedings of CICLing 2004*, Springer Verlag; pp. 215-228.
33. Hatzivassiloglou, V.; McKeown, K. R. Predicting the semantic orientation of adjectives. In *Proceedings of ACL/EACL-97*; 1997, pp. 174–181.
34. Turney, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02*; 2002, pp. 417–424.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2018 doi:10.20944/preprints201805.0102.v2

35. Purandare, A. Word Sense Discrimination by Clustering Similar Contexts. Ph.D. thesis, Department of Computer Science, University of Minnesota, August, 2004.
36. Dennis, S. A memory-based theory of verbal cognition. Cognitive Science 2005, 29, 145-193, DOI: 10.1207/s15516709cog0000_9
37. Sankoff, D.; Kruskal, J. B., eds. *Time warps, string edits and macromolecules: the theory and practice of sequence comparison;* Addison Wesley, 1983
38. Dennis, S. A comparison of statistical models for the extraction of lexical information from text corpora. In *Proceedings of the 25th Conference of the Cognitive Science Community*; 2003
39. Miller, G.; Charles, W. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 1991, 6(1), 1–28.
40. Redington, M.; Chater, N.; Finch, S. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* 1998, 22, 425-469.
41. Deerwester, S.; Dumais, S.; Landauer, T.; Furnas, G., Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 1990, 41(6), 391–407.
42. Hofmann, T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning* 2001, 42, 177–196
43. Brown, R.; Berko, J. Word association and the acquisition of grammar. *Child Development* 1960, 31, 1-14.
44. Alshahrani, S.; Kapetanios, E. Are Deep Learning Approaches Suitable for Natural Language Processing? In *21st International Conference on Applications of Natural Language to Information Systems (NLDB 2016)*; Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S., Eds.; Springer LNCS, Volume 9612, pp. 343-349, ISBN 978-3-319-41753-0.
45. Dong, L.; Wei, F.; Tan, C.; Tang, D.; Zhou, M.; Xu, K. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *Proc. ACL-2014*, 49–54.
46. Weston, J.; America, N. E. C. L.; Way, I. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proc. ICML* 2008, pp. 160-167.
47. Sun, Y.; Lin, L.; Tang, D.; Yang, N.; Ji, Z.; Wang, X. Modelling Mention, Context and Entity with Neural Networks for Entity Disambiguation. In *Proc. IJCAI*, 2015, pp. 1333–1339
48. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuglu, K.; Kuksa, P. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* 2011, 12, 2493–2537
49. Vector Representations of Words. Available online: https://www.tensorflow.org/tutorials/word2vec (Accessed on 30th April, 2018)
50. GloVe: Global Vectors for Word Representation. Available online: https://nlp.stanford.edu/projects/glove/ (Accessed on 30th April, 2018)
51. Ba, L.; Caurana, R. Do Deep Nets Really Need to be Deep ? In *arXiv preprint arXiv:1312.6184*; 2013, 521(7553), pp. 1-6
52. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, {ACL} 2014*; 2014, Baltimore, MD, USA, Volume 1, pp. 655–665
53. Santos, C.N. Dos; Guimarães, V. Boosting Named Entity Recognition with Neural Character Embeddings. In *Proc. ACL 2014*, pp. 25–33
54. Malinowski, M.; Rohrbach, M.; Fritz, M. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. *IEEE International Conference on Computer Vision*, 2015, 1-9
55. Johnson, R.; Zhang, T. Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding. In *Advances in Neural Information Processing Systems 28 (NIPS* 2015*)*, pp. 1-12
56. Irsoy, O.; Cardie, C. Opinion Mining with Deep Recurrent Neural Networks. In *Proc. EMNLP-2014*, pp. 720–728.
57. Jean, S.; Cho, K.; Memisevic, R. Bengio, Y. On using very large target vocabulary for neural machine translation. In *Proc. ACL-IJCNLP;* 2015
58. Shen, Y.; He, X.; Gao, J.; Deng, L.; Mesnil, G. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management - CIKM '14*; 2014, pp. 101-110
59. Wang, P.; Xu, J.; Xu, B.; Liu, C.; Zhang, H.; Wang, F.; Hao, H. Semantic Clustering and Convolutional Neural Network for Short Text Categorization. In *Proceedings ACL 2015*, pp. 352-357

doi:10.20944/preprints201805.0102.v2

60. Santos, C. N. Dos; Gatti, M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Coling-2014*; pp. 69–78
61. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. Available online: https://arxiv.org/pdf/1510.03820.pdf (Accessed on 2nd of May, 2018)
62. Griffiths, T.L.; Steyvers, M. Prediction and semantic association. *Advances in Neural Information Processing Systems* 2003, 15
63. Baroni, M.; Dinu, G.; Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. Annual Meeting of the Association of Computational Linguistics*; 2014
64. Roget, P. M. *Roget's International Thesaurus*, 7th ed.; Kipfer, B. A., Ed.; 2010
65. Miller, G. A. Wordnet: a dictionary browser. In *Proceedings of the First International Conference on Information in Data*; 1985, University of Waterloo, Waterloo.
66. Fellbaum, C. A semantic network of English: The mother of all wordnets. *Computers and the Humanities* 1998, 32, 209–220
67. Hamp, B.; Feldweg, H. GermaNet - a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*; 1997, Madrid
68. Burgess, C.; Kevin, L. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes* 1997, 12, 177–210
69. Rapp, R. The computation of word associations. In *Proceedings of COLING-02*; 2002, Taipei, Taiwan.
70. Jiang, J.; Conrath, D. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research on Computational Linguistics*; 1997, Taiwan.
71. Turney, P. D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of European Conference on Machine Learning*; 2001, pp. 491–502
72. Rothkopf, E.; Coke, E. U. Intralist Association Data for 99 words of the Kent-Rosanoff Word List. *Psychological Reports*, 1961, 8, 463-474
73. Nirenburg, S.; McShane, M.; Beale, S.; Wood, P.; Scassellati, B.; Magnin, O.; Roncone, A. Toward Human-Like Robot Learning. In *Proc. NLDB 2018, to appear*; 2018, Paris, France