Context Similarity and Semantic Relationships

Supervised Project

Bibliographic Report

Written by

Nami Akazawa and **Emre Canbazer**

Supervised by

Sylvain Pogodalla

MSc Natural Language Processing

Academic Year 2020-2021

Abstract

Nowadays, the use of vector representations of words (vector space models) is vital in the field of Natural Language Processing and Cognitive Science. Much active research relating to vector space models is ongoing, however, it is still an open question how introducing linguistic processing impacts the encoding of semantic relatedness in vector space models.

In this report, we first provide some contexts about the overall concept of word vector space models and main techniques used in building them as well as the previous works done by other researchers in building models especially the ones that incorporate lexical relation. In order for us to describe word meaning phenomena exist in language that are addressed in vector space models, we explain a linguistic theory. The theory introduces defined notions that give description and systematization of semantic relationships. We then present our approach of constructing vector space models with various syntactic relations used. The models can be analyzed whether there are any linguistic properties can be found and if there are, can this then be explained.

Contents

1 Introduction

There are various different methods for representing word as vector and various state-of-art distributional methods are applied to natural language processing (NLP) and cognitive science tasks such as automatic thesaurus extraction, information retrieval, and semantic priming to name few. These tasks benefit from a semantic vector space that can embed words and their features (such as the meaning of the word) and reflect relationship between words (e.g., similarities).

Our project aims to explore the lexical relations, if any, that can be captured by incorporating some linguistic properties to the distributional models. In particular, we are interested in using linguistic theories that explain and categorize semantic relatedness in languages.

This report aims to introduce the methodology of constructing and evaluating semantic space vector and presenting an introduction to some of the linguistic properties. In first part of Section 2, we first summarize the concept of distributional semantics and then explain the main techniques used in constructing distributional models of word meanings, as well as model parameters: similarity measure, weighting functions, and context definition. We focus on explaining two different distributional semantic methods: word-based models and syntax-based models. Since we want to relate context similarity obtained from distributional models to formal theories of semantic relatedness, we introduce various existing semantic relations. We describe in detail one linguistic theory, Meaning-Text Theory and especially it's explanatory combinatorial lexicology branch developed by Mel'čuk (2016), which provides a in-depth characterization of lexicographical definition and lexical functions. The last part of Section 3 explains existing evaluation methods for distributional models. In the conclusion, we outlines how we will proceed to the implementation phase based on what we learned from our investigations, as presented in this report.

It is in our interest that being able to find how lexical relationships are expressed in automatically constructed distributional models from textual data and make some observation that can contribute to understand more about the nature of the semantic relatedness exist in an embedded space.

2 Background

2.1 Distributional Semantics

The main factor that distinguishes distributional semantics from its predecessor, formal semantics, is the method that is adopted in the former. Formal semantics, by definition, attempts to explain the semantic phenomena by examining the constituents of a compound expression, having compositionality as method (Pulman, 2013), based on set-theoretic models (Clark, 2015). On the other hand, distributional semantics, rooting in the distributional hypothesis that is proposed by the structuralist linguist Zellig Harris (Turney and Pantel, 2010), prioritizes the distributional method and uses vector spaces for representing its models. Distributional method - with the term 'distribution' indicating the set of contexts in which the target word is observed to occur (Clark, 2015) - suggests that the meaning of a word is characterized by its surroundings, or its contexts, which can range from a few of words (to the right and to the left) to a paragraph or a whole document depending on the approach. To illustrate, Harris proposes a 'diagnostic frame' to determine the part-of-speech of a word, instead of relying on grammatical intuition (Pulman, 2013). Another example would be Firth's concept of collocations which has been important for computational linguistics since it emerged. What makes the Firthian notion of collocation special is its strict independence from compositionality and the fact that it prioritizes the environment of the target word (and not its interior structure) to explain its behaviour (Pulman, 2013), which serves to disambiguate the word meaning by taking into consideration its context. It is indeed this method that distributional semantics is based upon.

2.1.1 Vector Space Model

Given the idea of the distributional semantics and its hypothesis, every word can be represented as high-dimensional vectors in a common vector space. This vector space can encode the meanings of words and thus it can capture the semantic relation of words using any similarity or distance measures. The simplest construction of a vector space model (VSM) for words is that given a set of target words and corpus, we define a set of basis elements where it can be a collection of unique words, lemmas, words with their part-of-speech tag or dependency relation, and so on. The number of the dimensions for the semantic space will thus be the size of basis elements. Then, a target word's coordinates represent the frequency of each basis element occurring within a certain distance before and/or after the target word in the corpus.

To illustrate this definition, we give a simple example of word-word matrix (Figure 1). In this

example, given a set of sentences, we choose a context window size of four, meaning that only the frequency of four words to the left and four words to the right from a target word will be taken into consideration. We use all unique words in sentences (33 in total) as basis elements and four selected words as target, thus the created matrix does not take a consideration of the basis element words' syntactic relation to the target word and the word orders in each sentence. It simply keeps a collection of words, thus this model can be called as "bag-of-words" model. To give more explanation, the word *pie* appears two times in total within a context window for the target word *cherry* in the example sentences, thus the corresponding cell stores 2. On the other hand, if a basis element does not occur within a context window of a target word, then there will be 0 in the cell.

One of the variations of the VSM (Clark, 2015; Padró et al., 2014; Heylen et al., 2008; Peirsman, Heylen, and Speelman, 2007) is to look at the syntactic information of the target word and its surrounding text and only consider the ones with certain relations. VSM can also take morphemes, phrases including collocation, sentences, or documents instead of words as units to represent (Turney and Pantel, 2010). It is unclear, however, how the different context types relate to the overall quality and properties of the VSM (Baroni, Lenci, and Sahlgren, 2010).

In the later sections, we discuss in detail the two different approaches of building the VSM as well as the choices of distance measures that are relevant to our project.

Sentences:

Simple strawberry pie with fresh strawberries coated in a light strawberry glaze.
My favorite pie is cherry pie but I like apple pie as well.
A doctor opens the medicine cabinet to get drugs.
I go pharmacy to get medicine that I need.

Window size: 4
Target words: ⟨*strawberry, cherry, doctor, pharmacy*⟩
Basis elements: ⟨*simple, strawberry, pie, ..., my, favorite, ..., a, doctor, opens, ... go, pharmacy, that, need*⟩

	simple	...	pie	favorite	medicine	drugs
strawberry	1	...	1	0	0	0
cherry	0	...	2	1	0	0
doctor	0	...	0	0	1	0
pharmacy	0	...	0	0	1	0

Figure 1: A simple example of co-occurring space given a set of sentences from a text, showing five of the dimensions (for pedagogical purpose).

2.1.2 Weighting Schemes

Depending on the size of the target words and basis elements, the co-occurrence matrix can be large or small. However, counting the raw frequency of words' co-occurrence is very skewed and non-discriminative. If we consider all the unique words in text as basis elements, rare word pairs will be overly infrequent and it will result in sparseness (0s in many cells) of word-word matrix. One simple way to tackle this challenge is to have a frequency threshold to remove low frequency words. There also exists words such as ⟨*the, of, as, and*⟩ that usually have extremely high occurrences in the text, yet they seem to provide little information with the semantic relatedness of other words (e.g., "success","goal" and so on). One option to overcome this is the removal of such *stop words* from the basis elements.

What if we have an extremely dense and large matrix? We wish to weight more on highly frequent context words that are informative to the target words, and less on the words that are ubiquitous. In other words, we want the semantically related words to have higher correlation and those with no semantic relatedness to have lower values.

In following section, we introduce pointwise mutual information, a weighting algorithm that allows to eliminate word pairs that were falsely correlated from the matrix.

Positive Pointwise Mutual Information

Pointwise Mutual Information (PMI, Fano and Hawkins, 1961) is a measure of how often two events x and y occur together, compared with two events occurring independently. Given this definition, we can check if target word and context word co-occur more than if they were independent. The PMI between a target word w and a context word c is defined as following:

$$\text{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)} \tag{2.1}$$

The $\text{PMI}(w,c)$ allows us to quantify an estimate of how much more the two words co-occur in a window than we expect by pure chance. In the nominator, we compute the probability of how often we see two words w and c together by using maximum likelihood estimates (MLE). MLE estimates the probability of some word x by normalizing the number of observations for x, c_x by the total number of word tokens N:

$$\text{P}(x) = \frac{c_x}{N} \tag{2.2}$$

In our example, borrowing the example of Jurafsky and Martin (2020), let us assume that we have a co-occurrence matrix F with W rows of target words and C columns of contexts (basis elements), we can get the count of word w_i and c_i co-occurring by accessing f_{ij} cell. Applying this to MLE, we can

get $P(w_i, c_i)$ by:

$$\mathrm{P}(w_i, c_j) = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \tag{2.3}$$

The denominator is the multiplication of the individual distribution, the probability of w and the probability of c, telling us that how often we would expect for the independently occurring words w and c to occur together. Knowing that w appears in the text might also tell us something about the likelihood of c being present, and vice versa. By taking the ratio of these two, we can get an estimate of how much more the two words actually co-occur together than we expect by chance.

The value of $\mathrm{PMI}(w,c)$ falls in a range of negative to positive infinity. If PMI is positive, then (w,c) pair is more likely to occur together since $\frac{P(w,c)}{P(w)P(c)} > 1$ and thus $P(w,c) > P(w)P(c)$, implying that w and c occur mutually more than individually. On the other hand, a negative PMI value means that the two words are co-occurring less often than both of w and c or one of them occurring individually. Its negative value tend to be unreliable since it is unlikely to get many co-occurrences of a word pair in a limited size of text, or otherwise it shows uninformative co-occurrences, for example, 'the' and 'book'(where word 'the' is extremely used). Thus the suggested solution for this problem is to use Positive PMI (PPMI, Niwa and Nitta, 1994). PPMI replaces all negative PMI values with 0:

$$\mathrm{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0) \tag{2.4}$$

PMI is biased towards infrequent events. There are various ways to correct this bias empirically. One of them is to give rare words slightly higher probabilities (Jurafsky and Martin, 2020) . A slight modification to the computation of $P(c)$ to $P_\alpha(c)$ solves the problem:

$$\mathrm{PPMI}_\alpha(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P_\alpha(c)}, 0) \tag{2.5}$$

where

$$P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c count(c)^\alpha} \tag{2.6}$$

By raising the probability of the context words to the power of α (setting α to 0.75 has been found to be effective by Levy, Goldberg, and Dagan, 2015), the probability assigned to rare context words increases, and thus lowers their PMI scores.

Once the vector in semantic space is weighted, we can compute the similarity, distance or divergence between two words by using various similarity functions.

2.1.3 Distance Measures

Word pairs that have the highest similarity values (closest distance) computed by any distance measures are assumed to be semantically related. However, which measures are used is important as they can

produce different performance at the evaluation steps. Weeds, Weir, and McCarthy (2004) states that it is necessary to evaluate variety of distributional similarity measures in addition to evaluate a new VSM model or algorithm.

Padó and Lapata (2007) compares the impact of different similarity measures used to compute a word's distributionally closest neighbours. For their dependency-based model, they tested the following seven similarity measures: cosine, Euclidean distance, L_1 norm, Jaccard's coefficient, Kullback-Leibler divergence, skew divergence, and Lin's measure. Padó and Lapata (2007) report that among the other measures, Lin's similarity measure performed the highest Pearson product moment correlation ("Pearson's r") with human ratings. Weeds, Weir, and McCarthy (2004) conducted a study of the characteristic properties of different similarity measures used. They found that in a task of determining distributionally similar words to a target word, some similarity measures have a tendency to find words with a similar frequency as the target word. Interestingly, a measure that tends to select higher frequency words as distributionally similar neighbours performs significant results in the task of determining compositionality of collocations (Weeds, Weir, and McCarthy, 2004). One such measure is Jensen-Shannon Divergence which is a variation of Kullback-Leibler divergence. It is difficult to determine which similarity measures is best to use. Since our project aligns somewhat with the Padó and Lapata (2007)'s study, we follow their selections for the similarity measures. In this section, we discuss several similarity measures mentioned above.

Cosine Similarity

The dot product of two vectors is normalized by the division by the lengths of each of the two vectors. This is equal to the cosine of the angle between two vectors. Since the co-occurrence counts are non-negative, the range of the cosine for these vectors are in the range 0 to 1.

$$sim_{cos}(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{2.7}$$

Euclidean Distance

Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. The distance between word vectors w_1 and w_2 is defined as follows:

$$dist_{euc}(\vec{w_1},\vec{w_2}) = \sqrt{\sum_{i=1}^{n}(w_{1i} - w_{2i})^2} \tag{2.8}$$

L1 norm

L1 norm simply take the sum of absolute difference between the corresponding elements of the two word vectors.

$$Norm_{L_1}(\vec{w_1}, \vec{w_2}) = \sum_{i=1}^{n} |(w_{1i} - w_{2i})| \tag{2.9}$$

Jaccard's Coefficient

Jaccard's Coefficient measures how much two sets overlaps by taking the ratio between their intersection and their union. Its measure outputs the value ranging from 0 to 1.

$$coef_{ja}(w_1, w_2) = \frac{|Attr(w_1) \cap Attr(w_2)|}{|Attr(w_1) \cup Attr(w_2)|} \tag{2.10}$$

where $Attr(w)$ is the set of words $\{c_1, c_2, ..., c_k\}$ co-occurring with target word w.

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence, also known as the relative entropy, is a way to compare two probability distributions. We can measure the distance between the two distributions p and q by following:

$$D_{KL}(p||q) = \sum_{i=1}^{n} p \log \frac{p}{q} \tag{2.11}$$

where $p = P(c_i|w_1)$, $q = P(c_i|w_2)$ and w_1, w_2 are words.

However, the KL divergence is infinite when $p > 0$ and $q = 0$. It also has an asymmetric property where $D(p||q) \neq D(q||p)$. Thus, the direct use of KL divergence measure is avoided; instead, we use approximations of the KL divergence that do not require denominator q to be zero.

Skew Divergence

Skew divergence (Lee, 1999) is an approach where q distribution is smoothed using the p distribution.

$$D_{skew}(p||q) = D(p||(\alpha \cdot q + (1 - \alpha) \cdot p)) \tag{2.12}$$

where $p = P(c_i|w_1)$, $q = P(c_i|w_2)$ and w_1, w_2 are words. This measure avoids the indefinite problem that happens with the KL divergence by combining the two distributions p and q according to the parameter α.

Jensen-Shannon Divergence

Jensen-Shannon (JS) divergence (Lin, 1991) is another variant of the KL divergence that measures the distance of each distribution from their average.

$$D_{js}(p||q) = \frac{1}{2}\left(D\left(p||\frac{p+q}{2}\right) + D\left(q||\frac{p+q}{2}\right)\right) \tag{2.13}$$

where $p = P(c_i|w_1)$, $q = P(c_i|w_2)$ and w_1, w_2 are words. It uses the KL divergence to calculate a normalized score that is symmetrical, meaning $D_{js}(p||q) == D_{js}(q||p)$. It also circumvents KL divergence undefined problem by taking abstract means and will always having a finite value between 0 to 1.

Lin's measure

Lin (1998) has proposed a similarity theorem, which states, "the similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are." Thus two words are similar if their basis elements have similar information content.

$$sim_{Lin}(w_1, w_2) = \frac{\sum_{F(w_1)\cap F(w_2)}(I(c, w_1) + I(c, w_2))}{\sum_{F(w_1)} I(c, w_1) + \sum_{F(w_2)} I(c, w_2)} \tag{2.14}$$

where $I(c,w)$ is the PMI between c and w, and $F(w)$ is the set of basis elements c ($F(w) = \{c : I(c,w) > 0\}$) and

$$I(c, w) = \log \frac{P(c|w)}{P(c)} \tag{2.15}$$

2.1.4 Word-based Models

As we explained in the Section 2.1.1, traditional word-based co-occurrence models build their vector space by only considering a window of co-occurring words surrounding the target word. Researchers (Padó and Lapata, 2007; Heylen et al., 2008) usually use these methods for the comparison to their own models.

There are various possibilities regarding the definition of context. Kiela and Clark (2014) and Lison and Kutuzov (2017) study different parameter settings of the construction of VSMs. We discuss some of the specific aspects they investigate and their findings:

1. Vector size
2. Context window size
3. The relative position of the context window
4. Feature granularity

Vector size Commonly, dimension size is restricted to a relatively small number, and the dimension can simply be number of only k most frequent words (without the stop words) in a corpus. Kiela and Clark (2014) conclude with their experiment that 50,000 features or less produces generally good

performance. However, we are also interested in a qualitative interpretation of the semantic vector space, and depending on our context definition, a variety of different dimensions should be tested to find the optimal one.

Context window size Depending on the task, smaller or bigger window sizes work well. In summary, Kiela and Clark (2014) find that for the large corpora, small window size (3,5,7 and 9) works well. Again, depending on the context definition, setting to a larger window size or considering a whole sentence should be explored as well.

Relative position Paying equal attention to the words to the left and to the right of the target word is commonly used. Lison and Kutuzov (2017) experimented on considering only the left or right context words and interestingly, found that taking into account only context words to the right of the target window is as sufficient as the symmetric (considering right and left equally) option from the result of the spearman correlation.

Feature granularity Lemmatizing or stemming of the context words can help reduce data sparsity. Kiela and Clark (2014) test the impact of lemmatizing, stemming, and increase of granularity by pairing each context word with part-of-speech tag or a lexical category from CCG (Steedman, 2004). They found that stemming yielded the best overall performance, but that adding more granularity did not have any improvement.

2.1.5 Syntax-based Models

The abstract definition of context becomes sophisticated in the syntax-based models. The intuition of the syntax-based model is that we might be able to construct a semantically-enriched word vector space model that captures different semantic relations by incorporating information about the syntactic relationship between a target word and other words. Padó and Lapata (2007), Peirsman, Heylen, and Speelman (2007), and Heylen et al. (2008), each only consider the context words that satisfy a specific syntactic dependency relation to the target word.

Figure 2 shows an example of a matrix built by counting the co-occurrence of the syntactic relationship between words such as subject-verb and other relations. We can represent syntactic relations by tuple (r,w) where r is a relation type of a word w to a target word t. In this example, we use the basis vectors' term represented as (r,w). All the word-grammatical relation pairs in the example sentence constitute the basis vectors. We see a count of 1 in the cell of a row with a target word *ate* and a column of basis element *(subj, he)*, since in the example sentence, *he* is the subject of *ate*. Note that additionally, we can perform lemmatization. For example, the direct object of *ate* will correspond to the same basis vector of the direct object of *eat*. The idea is that by considering only the specific syntactic dependency relations, vector space model can be helpful to capture meaning of the target word.

Choices of what kind of syntactic relation one should use varies. Hagiwara, Ogawa, and Toyama (2008) have used indirect dependency in addition to normal direct dependency and shown the effectiveness in the acquisition of synonyms. In the study by Heylen et al. (2008), they consider eight syntactic relations (e.g., subject of verb, direct object of verb and modified by adjective). They find that their dependency model found more synonyms for high-frequency nouns and nouns that share semantic features of: object, event, property, situation, group, part, utterance, substance, location and thought. Padó and Lapata (2007) propose the use of *dependency paths* as contexts to build their vector space model. Given the dependency parse of the sentence, they define the context feature as *anchored* paths where the dependency starts at a particular target word. They only consider paths that have a maximum window size of k, that is, the absolute difference between the positions of the anchor (target) word and the context word with syntactic relation is at most k. They also discussed that some relations such as subjects and objects are more semantically informative than others. Thus they also constrained the context to be only a set of anchor paths with certain dependency relations. To quantify syntactic co-occurrence, they defined a path value function where the function takes into account the obliqueness hierarchy of grammatical relations. However there is frequency bias where words occurrence is not distributed evenly. To avoid words wrongly considered as similar due to their similar frequency, Padó and Lapata (2007) use a lexical association function to remove those randomly co-occurring contexts.

Clark (2015) raises a potential problem that using the dependency relations can result in data sparsity due to considering only the refined notions of the context. To overcome this, Heylen et al. (2008) simply remove the frequency cut off, to include all the relations that appears, whereas Padó and Lapata (2007) define a basis mapping function to map a feature (r, w) to just a word w as their final basis element.

Sentence:

He ate the cheese sandwich

Target words: $\langle$*he, ate, cheese, sandwich*$\rangle$
Basis elements: $\langle$*(subj, he), (root, ate), (det, the), (mod, cheese), (obj, sandwich)*$\rangle$

	(subj, he)	(root, ate)	(det, the)	(mod, cheese)	(obj, sandwich)
he	0	0	0	0	0
ate	1	0	0	0	1
cheese	0	0	0	0	0
sandwich	0	0	1	1	0

Figure 2: A simple example of Lin's (Lin, 1998) syntax-based semantic space.

2.2 Semantic Relations

VSM can be applied to extract different semantic properties and in this section, we explain those relationships.

Syntagmatics and paradigmatics are developed by pioneering structuralists such as Saussure, Jacobson, and Hjemslev so as to distinguish two kinds of linguistic phenomena: the former concerns combination and the latter concerns substitution. Paradigmatic relations are extensively used in dictionaries and other knowledge organization systems, whereas syntagmatic relations are mainly associated with co-occurrences of lexical units (Chiu and Lu, 2015). Paradigmatics indicates a process of lexical selection, e.g., what word to employ in the sentence and syntagmatics point out the arrangement of this 'chain of words' to form the meaning. It is also generally accepted in the field of linguistics that paradigmatic relation is implemented along the vertical axis and similarly the syntagmatic relation is implemented on the horizontal one. For instance, consider the sentence *The cat ate the mouse.*, commonly given example to illustrate these concepts. The lexical unit EAT is syntagmatically related to CAT, MOUSE and THE. Moreover, every word in the sentence could be replaced by any word of the same word class (and not necessarily by a synonym) (Finch, 2000). For instance, THE could be replaced by A, CAT could be replaced by DOG and EAT could be substituted by LOVE etc. With these two relations, numberless sentences can be formed.

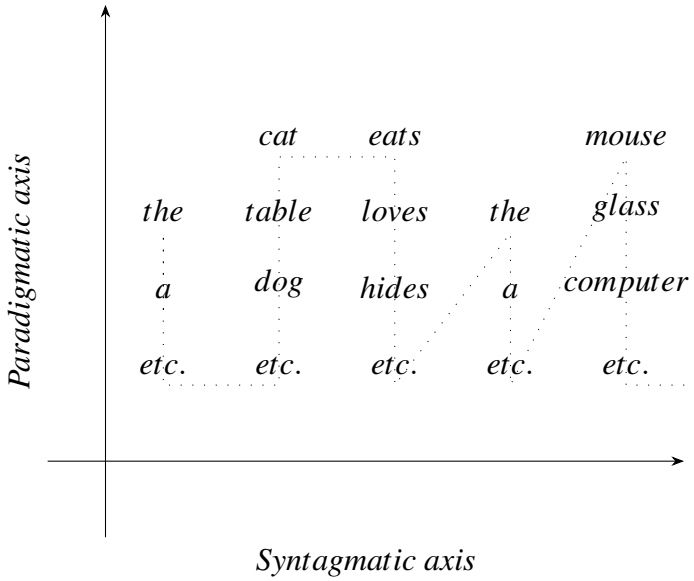

Figure 3: A simplified example of lexical selection (y axis) and syntactic co-occurrence (x axis).

2.2.1 Syntagmatic Relation

Without a doubt, it is an exceedingly difficult challenge for foreign speakers of a particular language (and even, to some extent, native ones) to get accustomed to using the words with correct pairs. Even when one chooses the right word that goes perfectly with what they need to say, it is possible to use it with improper ones and end up with a speech that does not sound 'natural' to the ear. The problem that is experienced in this case is connected to syntagmatics. Syntagmatically related words are ones that are likely to co-occur in the same text region. To illustrate: in English, the proper verb to use with EXAM is *to take*, however in French, one *passes* an exam (*passer* un EXAMEN) regardless of the result of the exam being successful or not, and in Turkish, one *enters* an exam (SINAVa *girmek*). Or, in English one *flies into* rage, but in French, one *puts themselves into* rage (*se mettre en* COLÈRE and in Russian, one *falls into* rage (*vpadat* RAŽ) (Mel'čuk, 2016). These phraseological expressions are called collocations and are gaining more and more importance in the description of linguistic phenomena.

Besides, Mel'čuk explains linguistic dependency through syntagmatics, which is crucial for our work since dependency-based models are widely discussed within the VSM literature. In a phraseme, one wordform necessarily dominates the other. This hierarchy is one-way and denoted as $w_1 \rightarrow w_2$ (Mel'čuk, 2009). In other words, it is similar to the logical operation of implication; w_1 implies w_2, thus w_2 depends on w_1. Since our work is focused on semantic relationships, the kind of dependency that we (briefly) cover here is limited to semantic dependency. The criterion of semantic dependency is predicativeness; the governor (w_1) is the predicate that takes w_2 as an argument. To illustrate, in the sentence *Mary wants to kiss John*, KISS depends on MARY (Mel'čuk, 2016).

2.2.2 Paradigmatic Relation

Words that are paradigmatically related (paradigms) are lexical units that occupy the same position in the sentence (Chomsky, 2006). Thus, they cannot co-occur; the speaker is obliged to make their decision to use the correct one. In other words, they are in opposition to one another (Mel'čuk, 2016). Returning to the sentence 'The cat ate the mouse.', the lexical unit THE, the one that depends on CAT, as we already mentioned above, could be replaced by A. So we could change the sentence into 'A cat ate the mouse' but not into *'A the cat ate the mouse.' nor into *'The a cat ate the mouse.' (In linguistics, unacceptable sentences are conventionally marked with an asterisk when they are used for demonstration purposes.). Because their role is the same. Paradigms are not necessarily synonyms, but synonymy is an excellent example for paradigmatically related lexical units since when we, for instance, choose to use TRANSPARENT in 'The water is transparent.', we agree to attribute SEE-THROUGH to WATER. Antonymy works in the same way in this case. We cannot attribute both OPAQUE and TRANSPARENT to WATER in the sentence in question, giving them the same position. Other than these two bidirectional paradigmatic relations, we also have hyponymy/hyperonymy and meronymy/holonymy. Hyponymy is the 'kind of' relation (Murphy, 2003), this means that if X is a hyponym of Y, then X is a type of

Y. Oppositely, Y is a hyperonym of X. They have a paradigmatic relationship since they can be used interchangeably at the same position in a sentence (e.g "The LIQUID is transparent."). Aside from these taxonomic relations, another type of semantic relation that is worth mentioning here is the part-whole relation, where meronymy and holonymy get on the stage with the holonym being the whole and the meronym being the part of it (Cruse, 2011). For instance, since WATER is a part of OCEAN, creating a part-whole relation, our sentence could change to "The ocean is transparent.". These relations are the ones that one could observe (or hopes to observe) on the lists related word lists that are extracted using VSMs.

2.2.3 Explanatory Combinatorial Lexicology

Meaning-Text Theory

It would be reasonable to think that, in order to mention Explanatory Combinatorial Lexicology, we must first introduce the Meaning-Text Theory (MTT), which frames this type of lexicology.

Theorized by Soviet linguists Aleksandr Žolkovskij and Igor Mel'čuk in the 1960s (Kahane, 1984), the MTT describes the natural language as a system of correspondences between linguistic meanings (the 'thing' that is intended to express) and linguistic texts (the utterance corresponding to the linguistic meaning) (Mel'čuk, 2016). According to Mel'čuk, there are two types of linguistic operations: The first one is performed by the Speaker, who associates in his head all the texts that may be utilized to express the intended meaning, e.g., the set of all the synonyms known to the Speaker, and chooses the right one that corresponds to the certain context. This operation is called *linguistic synthesis* and according to the MTT, it comes first in speech activity. The second operation is *linguistic analysis*. Performed by the interlocutor, or Addressee in Mel'čukian terminology, this postlinguistic operation consists of associating all the meanings that the chosen text may signify, e.g., the set of all the homonyms of the text, and selecting the right one. The former is prioritized by the MTT whereas the latter is described as optional for the modeling of natural language.

MTT accepts the idea that the correspondence between the meaning and the text is many-to-many, which means that several meanings correspond to several texts and vice versa. This aspect of the theory could be put into words as the following expression:

$$\{SemR_i\} \iff \{PhonetR_j\} \mid i, j > 0$$

In the expression above, $\{SemRi\}$ indicates the semantic representation of the meaning i and $\{PhonetRj\}$ indicates the phonetic representation of the text j.

The Role of Lexicon in MTT

Within the MTT, the lexicon of a language *L* is, along with grammar, one of the two components of the language. It consists of the lexical units (LUs) of *L* (Mel'čuk, 2016). The LU identifies a correspondence between the signified and the signifier (the two composers of a linguistic sign in the Saussurian sense).

Although it is only one of the two components of the language, the lexicon is prioritized within the Meaning-Text approach whereas grammar is considered as just a system of rules on how to put together the LUs. The lexicon of a language *L* is documented in the form of an Explanatory-Combinatorial Dictionary (ECD) (Mel'čuk, 2016).

Towards an Explanatory-Combinatorial Lexicology

A proto-ECD was constructed for the Russian language in 1984 by the first explanatory-combinatorial lexicographers, Mel'čuk and Žolkovskij to provide a scientific description of the language. This dictionary was designed to be exceptionally versatile; it would serve not only as a dictionary to look up words but also as a linguistic work to model a language, as a pedagogical manual for language learners, as a source of reference for translators, editors ad journalists and most importantly for our field of study, as a lexical database for automatic text processing (Mel'čuk, 2016). Mel'čuk proposes three factors that distinguish the ECD from traditional dictionaries; its theoretical orientation, its connection to a systematic linguistic theory, namely the Meaning-Text theory, its active character which not only allows one to analyze the text but also to produce (synthesize, in Mel'čukian terms) the text, and lastly its formalization which consists of a metalanguage specific to the ECD that enables the user to comprehend the entry without any extralinguistic knowledge.

A lexical entry in an ECD is composed of three parts: the semantic zone, the syntactic combinatorial zone, and the lexical combinatorial zone (Mel'čuk, 2016). The semantic zone is the part of the entry that includes the lexicographic definition of the LU. However, the definition in ECF is not a traditional one, it explains the LU with the help of *semantic actants*, which are variables that we can consider as the augmented version of *sb* and *sth* in a traditional dictionary. To illustrate: so as to explain the lexeme ALLOW, we introduce 'X allows Y to do Z' in this zone and then go into detail about the role of each actant in the next one. Moreover, the definition incorporates a semantic decomposition, namely the procedure of splitting the complex LUs into simpler semantemes (= semantic units). For example, the LU KILL would be expressed as *X kills Y* which is semantically decomposed into an approximation as *X, by acting upon Y, causes that Y dies* (Mel'čuk, 2009).

The syntactic co-occurrence zone introduces the **government pattern** (GP) of the LU. In tabular form, the GP serves to demonstrate the possible uses of the LU in a sentence in a formalized way, e.g., without having to illustrate it in a sentence. For example, the GP notation for the LU KNOW, or *X knows Y* is as follows (Mel'čuk, 2015):

X $\Longleftrightarrow$ *I*	Y $\Longleftrightarrow$ *II*
N	1. N
	2. *that* CLAUSE
	3. *wh-* CLAUSE
	4. *whether/if* CLAUSE

Figure 4: A simplified example of a GP in tabular form.

The GP above tells us that the actant X is implemented by a noun and the actant Y is implemented either by a noun, or by a subordinate clause followed by 'that', or by a subordinate clause followed by a 'wh-' word such as 'what', 'when', 'who' etc., or else by a subordinate clause followed by 'whether' or 'if'.

And finally, the third zone of the lexical entry of an ECD, the semantic derivation and lexical cooccurrence zone, is the one that is reserved for **lexical functions**. Through lexical functions, this zone demonstrates by what other variety of the LU in question could be replaced (the paradigmatic aspect) and also with what other LUs it is likely to co-occur (the syntagmatic aspect). When called with a lexical unit L as an argument, the function f evaluates to L', this is notated as $f(L) = L'$. For instance, if a paradigmatic lexical function f_i indicates 'someone who does...' then f_i(CRIME) gives CRIMINAL and analogically f_i(LECTURE) evaluates to LECTURER (Mel'čuk, 2016). This is the *semantic derivation* part of this last zone. The second part constitutes a syntagmatic operation; if the function f_j is a syntagmatic lexical function that has the meaning 'do...' then f_j(CRIME) gives us COMMIT and similarly f_j(LECTURE) evaluates to DELIVER since this word is the one that collocates with the word 'lecture' in this sense.

2.3 Evaluation

Once we have VSMs, we then need to quantitatively evaluate the quality of them on different tasks. There are two types of evaluation methods, intrinsic and extrinsic. It might be necessary to perform both evaluation methods since the model performance in the intrinsic methods might not have similar correlation with the results of the extrinsic methods. However, the choice of which method to use really depends on the research goals (Padó and Lapata, 2007). Here we discuss several tasks that are used by related research.

2.3.1 Intrinsic Tasks

Intrinsic evaluation methods involve the comparison between the manually collected gold data (e.g., thesauri) and the semantic space on specific intermediate sub-tasks (such as analogy completion and

semantic similarity check). One downside of this evaluation method is that the automatically extracted VSM might capture correct semantic relations for some target word that are not listed in the manually created golden data. Also, the reliability of the gold data can be questionable. A common measure for this evaluation is **precision at rank** k, that is a proportion of recommended words in the top-k set for a target word that are relevant. A total score is calculated by the sum of each target word's total number of how many predicted synonym words actually match the words in the "gold standard" thesaurus, and then average it by the number of target words.

Heylen et al. (2008) performed two intrinsic evaluations using the Dutch part of EuroWord-Net (Vossen, 1998) dataset as a gold standard. They analysed the overall performance of the VSMs by measuring the average semantic similarity of the nearest neighbours to their target words as recorded in EuroWordNet. They also evaluated if a word computed as a nearest neighbour falls into four semantic relations they defined (synonym, hyponym, hypernym and cohyponym) with its target according to EuroWordNet. Their dependency-based model found a tightly related neighbour for 50% of the target words and a true synonym for 14% and outperformed other word-based models.

2.3.2 Extrinsic Tasks

Extrinsic evaluation of word vectors is the evaluation of a set of generated word vectors on the real task at hand. Various domain tasks are applied by researchers. For example, Padó and Lapata (2007) use a psycholinguistic task, *semantic priming* (Lowe and McDonald, 2000) to evaluate their dependency models. Semantic priming describes the observation that a human response to a target word (e.g., car) is quicker when it is preceded by a semantically related word that is called as prime (e.g., truck) compared to an unrelated word (e.g., apple). They especially focused on single-word lexical priming study performed by Hodgson (1991). He found the priming effects that the reading times for target-prime pairs are reduced for several types of semantic relation, including category coordinates, synonyms, conceptual relation (e.g., election-vote), phrasal relations (e.g., private-property) etc. Given this study, Padó and Lapata (2007) model the reading time for prime-target pairs using the semantic similarity between the prime and target described as vectors. They found that their syntax-based model resulted higher correlation than the word-based models.

Padó and Lapata (2007) also perform *synonymy detection* using TOEFL (Teaching of English as a Foreign Language) tests set consisting multiple-choice questions. The task required identifying the closest synonym of a given target word from 4 potential synonym words. An example question given by Padó and Lapata (2007) is following:

You will find the office at the main **intersection**.
(a) place (b) crossroads (c) roundabout (d) building

In this example, the closest synonym to target word **intersection** is **crossroads**. The method is to select a word predicted as closest synonym to the target in the automatically constructed VSM. Again,

Padó and Lapata (2007) find that their syntax-based model (accuracy 73.0%) outperformed the word-based model (accuracy 61.3%).

As a third task, Padó and Lapata (2007) carried out *word sense ranking* in context. One crucial requirement to accomplish many NLP tasks is the ability of determining which "sense" (meaning) of a word is meant by the use of the word in a particular context. They followed the assumption that a sense of a target word can be determined by its distributionally similar neighbors' senses. As the result, their dependency model had higher performance than the word-based model.

3 Conclusion and Future Works

From this report, we learn that semantic vector space is constructed based upon on the hypothesis that the semantic similarity of words can be predicted from their embedded distributional similarity. However, there are questions about how distributional models encode such semantic information about words and what kind of semantic relations (e.g., synonymy) can be captured. Our aim is to investigate the statistical and linguistic properties of sets of distributionally similar words returned by different parameter settings we have explained in Section 2. We will try to relate these properties to linguistic theories, particularly, to explanatory combinatorial lexicology and ideally classify the semantic relations between collocates automatically.

This report has shown that there are many parameter settings that could be used to explore and observe different outcomes possibly seen in VSM. By now, readers should also have some understanding of the existing semantic relations and a linguistic theory. For the next step, we will familiarize ourselves with different existing frameworks, especially MANGOES software [1] and experiment using it. MANGOES is a toolbox for constructing and evaluating word vector representations. We will use it to explore different state-of-the art unsupervised methods and evaluate them to understand the general flow of the experiment of VSMs.

We then will be constructing VSMs that incorporate syntactic relation that could capture semantic relation from textual data and study how they are expressed in VSMs. Starting by choosing some semantic relations that interest us, we will then try out different possibilities combination that are available in vector size, context window size and its relative position, weighting function, similarity measures, and syntactic relations. A significant part of our work will be finding the best performing model to carry all other subsequent experiments. Hence, it will require us to carefully define the parameter settings. Then we will evaluate our VSMs to different tasks and assess if there are some meaningful relations in our VSM and context similarity to formal theories of semantic relatedness. Additionally, it would be easy to understand if we could visualize our VSMs.

[1] link: https://gitlab.inria.fr/magnet/mangoes/-/tree/master

References

Baroni, M., Lenci, A., and Sahlgren, M. (2010). "Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models. Beyond Words and Documents". In:

Chiu, W. and Lu, K. (2015). Paradigmatic relations and syntagmatic relations: How are they related? *Proceedings of the Association for Information Science and Technology* [online]. 52.1, pp. 1–4. DOI: https://doi.org/10.1002/pra2.2015.1450520100122.

Chomsky, N. (2006). *Language and Mind*. 3rd ed. Cambridge University Press. DOI: 10.1017/CBO9780511791222.

Clark, S. (2015). "Vector Space Models of Lexical Meaning". In: *Handbook of Contemporary Semantics, 2nd edition.*

Cruse, A. (2011). *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford University Press UK.

Fano, R. and Hawkins, D. (1961). Transmission of Information: A Statistical Theory of Communications. *American Journal of Physics*. 29, pp. 793–794.

Finch, G. (2000). *Linguistic Terms and Concepts*. Palgrave. DOI: 10.1007/978-1-349-27748-3.

Hagiwara, M., Ogawa, Y., and Toyama, K. (2008). Effective Use of Indirect Dependency for Distributional Similarity. *Journal of Information Processing*. 15, pp. 19–42.

Heylen, K., Peirsman, Y., Geeraerts, D., and Speelman, D. (2008). "Modelling Word Similarity: an Evaluation of Automatic Synonymy Extraction Algorithms". In: *LREC*.

Hodgson, J. M. (1991). Informational constraints on pre-lexical priming. *Language and Cognitive Processes*. 6, pp. 169–205.

Jurafsky, D. and Martin, J. H. (2020). *Speech and Language Processing (3rd ed. draft)*.

Kahane, S. (1984). The Meaning-Text Theory. In: *Dependency and Valency, Handbooks of Linguistics and Communication Sciences*. De Gruyter.

Kiela, D. and Clark, S. (2014). "A Systematic Study of Semantic Vector Space Model Parameters". In:

Lee, L. (1999). Measures of Distributional Similarity. *ArXiv*. cs.CL/0001012.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*. 3, pp. 211–225.

Lin, D. (1998). "An Information-Theoretic Definition of Similarity". In: *ICML*.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*. 37, pp. 145–151.

Lison, P. and Kutuzov, A. (2017). Redefining Context Windows for Word Embedding Models: An Experimental Study. *ArXiv*. abs/1704.05781.

Lowe, W. and McDonald, S. (2000). "The Direct Route: Mediated Priming in Semantic Space". In:

Mel'čuk, I. (2009). Dependency in Natural anguage. In: *Dependency in Linguistic Description*. John Benjamins, pp. 1–110. DOI: 10.1075/slcs.111.03mel.

Mel'čuk, I. (2015). *Semantics: From Meaning to Text, Volume 3*. John Benjamins. DOI: https://doi.org/10.1075/slcs.168.

Mel'čuk, I. (2016). *Language: From Meaning to Text*. Academic Studies Press.

Murphy, M. (2003). *Semantic relations and the lexicon : antonymy, synonymy, and other paradigms*. Cambridge University Press.

Niwa, Y. and Nitta, Y. (1994). "Co-Occurrence Vectors From Corpora Vs. Distance Vectors From Dictionaries". In: *COLING*.

Padó, S. and Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*. 33, pp. 161–199.

Padró, M., Idiart, M., Villavicencio, A., and Ramisch, C. (2014). "Comparing Similarity Measures for Distributional Thesauri". In: *LREC*.

Peirsman, Y., Heylen, K., and Speelman, D. (2007). "Finding semantically related words in Dutch: co-occurrences versus syntactic contexts". In:

Pulman, S. (2013). Distributional Semantic Models. In: *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*. Ed. by C. Heunen, M. Sadrzadeh, and E. Grefenstette. Oxford University Press, ISBN 978-0-19-964629-6, pp. 333–358.

Steedman, M. (2004). "The syntactic process". In: *Language, speech, and communication*.

Turney, P. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* [online]. 37, pp. 141–188. DOI: 10.1613/jair.2934.

Vossen, P. (1998). "EuroWordNet: A multilingual database with lexical semantic networks". In: *Springer Netherlands*.

Weeds, J., Weir, D. J., and McCarthy, D. (2004). "Characterising Measures of Lexical Distributional Similarity". In: *COLING*.