A Survey of Linking Techniques for Neuroimaging Data and Computational Linguistics Models

Amir Bakarov
National Research University
Higher School of Economics,
Moscow, Russia
amirbakarov at gmail.com

Artyom Stepanov
National Research University
Higher School of Economics,
Moscow, Russia
stepartm@gmail.com

Anastasia Yashchenko
National Research University
Higher School of Economics,
Moscow, Russia

Abstract

This survey provides a comprehensive overview of the research, related to linguistic information processing in the human brain. First, we introduce the popular techniques, used for natural language processing. Next, we provide an information on methods of neurovisualization and their peculiarities. Finally, we summarize the results of the research, where the aforementioned techniques are used to explore how the human brain reacts to various linguistic items(e.g. words, texts etc).

1 Introduction

According to an ubiquitous schema, human language has five main levels: phonetics, morphology, semantics, syntax and pragmatics. Computational modeling of most of these levels gained a decent success in recent years with a help of artificial neural networks and large annotated resources. However, it is still not super good for semantics since theory of meaning always was one of the most tough stumbling stones for linguistics.

Approaches to semantics are very heterogenous and different by the nature. One of the most influential theories of meaning in philosophy and linguistics is the theory of reference, as formalized in set-theoretic semantics. The core proposal of set theory is that words denote things in the world Frege (1892); Tarski (1944); Montague (1973). So the word *cat*, for instance, has a so-called *extension* which is the set of all cats in some world. Set theory is closely related to truth theory in that it is possible to compute the truth or falsity of a statement for a particular world just by looking at the extension of that statement in that world. The basic notion of extension is complemented by the concept of "intension" which, under the standard account, is a mapping from possible worlds to extensions, i.e. a function which, given a word, returns the things denoted by that word in a particular world. Intension allows us to make sense of the fact that Evening Star and Morning Star have different connotations, although they denote the same object in the world: they simply have different intensions.

On the other hand, there are semantic theories that consider meaning comes from engaging in a set of normative human practices, so, in other words, semantics emerges from pragmatics being anchored to contexts. The linguistic theory of meaning closest to Wittgenstein's line of argumentation is distributionalism. In this approach, the meaning of cat is not directly linked to real cats but rather to the way people talk about cats. The collective of language users acts as a normative force

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

by restricting meaning to a set of uses appropriate in certain pragmatic situations. The roots of distributionalism can perhaps be found in Bloomfield (1936), but the theory grew to have much influence in the 1950s Harris (1954); Firth (1957). Some time later, in the 1990s, the advent of very large corpora and the increase in available computing power made the claims (to some extent) testable.

This theory makes meaning boundaries much less clear than they are in set theory, but one of the most cool things about distributional semantics is that it could be applied to computational semantic models as well as be integrated with psychological and cognitive theories of semantics. Both psychologists and linguists started investigating the idea that a word's meaning could be derived from its observed occurrences in text Landauer and Dumais (1997); Grefenstette (2012); Schütze (1998). These empirical efforts would soon lead to a very active area of research in computational linguistics called distributional semantics: a field which attempts to model lexical phenomena using distributions, i.e. patterns of word usage in large corpora.

Actually, exactly the distributional approach gained a popularity in the task of modeling language semantics. Different algorithms based on very different techniques Landauer and Dumais (1997); Mikolov et al. (2013), but exploiting the same distributional hypothesis, helped in resolving a broad range of natural language processing tasks, ranging from sentence boundary detection and named entity recognition to information retrieval and question answering.

However, such models easily fail and have lot of different limitations, being just an approximation of the notion of semantics from the perspective of true semantics. Actually, we do not know *what true semantics is* considering true semantics as an underlying semantic structure in a human cognition. Some psycholinguistics theories consider cognitive concept meaning as a graph, ...

There is no any extensive work surveying attempts to find an interplay between computational semantic models and the results of empirical experiments on neural and cognitive semantic processing. Our work is brought to cover this lacuna, helping future researcher to understand the context of previous studies in the perspective field of distributional cognitive semantics.

This paper is organized as follows. Section 2 is dedicated to a brief survey of current advances in computational semantic models, particularly, in different approaches to distributional hypothesis. Section 3 overviews current techniques of empirical cognitive experiments, while Section 4 actually brought to highlight recent works on an intersection of two these fields. Section 5 considers future challenges and prospective frontiers on the border of distributional semantics and cognitive studies.

2 Computational models of meaning

2.1 Distributional Semantic Models

Distributional semantic models exploit distributional theory based on corpus statistics through a concept of vector spaces. Words contexts are embedded into real-valued vectors, proposing all possible arithmetic operations in a vector space therefore giving ability to operate with words meaning: for instance, one of the most popular uses of distributional semantic models is to compute a degree of synonymy between two words. The key point here is to obtain from the corpus the most 'correct' vector representations of contexts.

The first approach to capturing contexts is based on the concept of *word-context matrix*. In this matrix, each unique word in a corpus should associated with a row columns of all words that can be encountered in a context of this word. So each cell in this matrix will have a certain value reporting degree of closeness of word a to word b from the context perspective.

However, in recent years a breakthrough in distributional semantics made most ubiquitous a second class of models which are prediction based.

2.1.1 Count-based models

1. **Mutual-information Weighted Word Co-occurrence Matrices.** Exploits Pointwise Mutual Information score (or Positive Pointwise Mutual Information)
2. **Singular Value Decomposition on Term-document Matrix** (Latent Semantic Analysis, Latent Semantic Indexing). The notions of Latent Semantic Analysis and Latent Semantic

Indexing are actually interchangeable, and the only difference is that they came from different fields of science. Latent Semantic Indexing was deployed in information retrieval systems in early 90s, while Latent Semantic Analysis came slightly later, being used for research of cognition and language, particularly exploring psycholinguistic models of human lexical acquisition.

3. **Brown Clusters**. Quite 'oldschool' technique with name origining from Brown corpus. The idea is to split a corpus to a set of clusters, initially association each word with its own one and then merging pairs of clusters.

2.1.2 Prediction-based models

1. **Classic Neural Language Models.** The first prediction-based distributional semantic model came in 2003 set up the trend of using neural language models as a training algorithm for distributional word representations Bengio et al. (2003). This language model was simple, consisting of a one-hidden layer feed-forward neural network with a loss trying to maximize probability of encountering the next word in a sequence.
2. **SENNA Embeddings.** Collobert and Weston (2008); Collobert et al. (2011) Until 2008, the research in distributional semantic models was not effective since computational powers were not able to give ability to train a model on a corpus of a decent size able to model word semantics. But in 2008 a classic Bengio's model was modernized with a killer feature: instead of maximizing the probability of the next word given the previous words (proposing cross-entropy loss that leads to a high computational complexity), the new network tried to output a higher score for a correct word sequence than for an incorrect one Collobert and Weston (2008). That allowed researchers and engineers to train first decent neural-based distributional semantic models, but the time and effectiveness of training still was pretty low.
3. **Word2Vec.** Mikolov et al. (2013) Word2Vec came out in 2013 was actually a revolution in a field of distributional semantic models that drew a huge popularity to this field. In contrast to Bengio's model and the Collobert model that can only base their predictions on past words, Mikolov's model aims to predict each next word in the corpus. Actually Word2Vec is not a name of a single neural network, but it is a common name of two different architectures: Continuous Bag of Words that predicts words on a base of their contexts, and Skip-Gram that vice verse predicts contexts on a base of encountered words.
4. **GloVe.** Pennington et al. (2014) GloVe was the first neural distributional model in which production of word vector representations was the main focus. The idea of a neural network based on counting ratios of the co-occurrence probabilities of two words instead of their co-occurrence probabilities themselves.

3 Neuroimaging in linguistic studies

3.1 Neuroimaging methods

In attempt to understand human nature, we try to understand how our brains works. Modern techniques and recent discoveries in neuroimaging and neurovisualization allow scientists to achieve great results in this king of study. In order to review the latest achievements in the field, we present here the most outstanding techniques and related work.

3.1.1 Electroencephalography

One of widely occuring neuroimaging method is EEG (Electroencephalography), which shows even the slightest changes in human cortex, measuring voltage fluctuations resulting from ionic current within the neurons of the brain. The main advantage of this technique is high time resolution (the shortest period of neural activity registered by EEG), that is why core measurements are made in special scale of ERP - event-related potentials, the most important moments of brain activity. One could call the N400 (400 milliseconds after stimuli were presented) the most important one for a linguist, because it is associated with comprehension of semantic information. Also N1-moment (80-120 milliseconds) is very essential as the earliest reaction of human brain.

This method is often used in experiments with brain-lesioned patients: in Paulmann (2010) means of EEG revealed that processing of emotional speech differs in cases of brain-lesioned individuals and healthy ones. After 200 milliseconds of normal processing, patients' brains showed that lesion of a specific region might impair some ERP components but not others (for more information see Cyma Van Petten (1990)).

Another study Armando Freitas da Rocha and Alfredo Pereira (2015) approves distributional nature of language processing by analysis of temporal and spatial characteristics of neural activation. Using Principal Component Analysis authors present 4 different models, describing activation of neurons sets, activated during processing of written and oral texts.

More information about using EEG in linguistic research one could find in this review: Lesya Y. Ganushchak (2011)

3.1.2 Magnetoencephalography

Very similar technique is MEG(Magnetoencephalography). During MEG-measurements we also observe natural voltage fluctuations, but using magnetometers. Both these techniques have very high temporal resolution (here we mean sub-millisecond scale), which makes them capable to register the slightest reactions of human brain. Unfortunately, both EEG and MEG suffer from inaccurate reaction localization due to space capability only till 5mm. Thus, these techniques are mostly used in such research,where the most essential aim is to investigate speed or intensity of brain function.

Usage of MEG helped to investigate phonological issues in cases of individuals with autism spectrum disorders in Brennan (2016). During listening to stimuli with legal and illegal phonological patterns,children (8-12 years old) brain activation was registered by MEG technology(around 330ms after the onset of the critical phoneme). Collected data reveals that respond to illegal stimuli was much more attenuated within ASD group then within control group what leads us to approving the idea of initial acoustic processing.

3.1.3 Positron Emission Tomography

Not so popular, as previous two, but undoubtedly useful in specific research, positron emission tomography (PET) is a technique that measures physiological function. It catches a set of characteristics showing activation of cognitive processes, such as blood flow, metabolism, neurotransmitters and others. PET is based on the detection of radioactivity emitted after a small amount of injected radioactive tracer. Nature of this method (invasive) may be the cause of its low frequency of usage, although the total radioactive dose is similar to the dose used in computed tomography.

Mostly it is used for medical conditions' research (for example, in understanding strokes and dementia), but it may be also used in linguistics research. Using lexical decision task and semantic categorization, scientists tried to find differences in processing of two word classes - nouns and verbs - in Collette (2001).

3.1.4 Magnetic Resonance Imaging

Another neuroimaging method MRI became wide-spread technique in neuroimaging due to capability of making extremely precise and detailed maps of human brain. Although MRI has not a remarkable merit of high temporal resolution as MEG or EEG , space accuracy, low invasivity and lack of radiation makes this method one of the most convenient for researchers. This capability to create detailed maps of human brain causes necessity of new means to measure brain space. Scientists use voxels - three-dimensional rectangular parallelepipeds, the size of which is determined by the thickness of the slice, the slice area, and a grid superimposed on the slice by scanning and depends on terms of a particular experiment. Larger voxels are used in full brain research, while in those that specialize in specific regions it is favourable to use smaller voxels. Size of a voxel may vary from 5 to 1 millimeters, providing high spatial resolution. But the slow response time of the registered brain activation limits its capability to characterize how, rather than where, the brain performs its cognitive processes.

Special mapping paradigm,called the activation likelihood estimate (ALE) technique, is used to navigate through MRI-data. These activations formed a left-lateralized network, including 7 regions: posterior inferior parietal lobe, middle temporal gyrus, fusiform and parahippocampal gyri, dorsomedial prefrontal cortex, inferior frontal gyrus, ventromedial prefrontal cortex, and posterior cingulate gyrus. The cortical regions involved in semantic processing can be grouped into 3 broad categories:

posterior multimodal and heteromodal association cortex, heteromodal prefrontal cortex, and medial limbic regions.

Secondary analyses showed specific subregions of this network associated with processing of a particular kind of information, e.g. abstract concepts or actions. These results were achieved by many linguistics studies, the most representative of which are briefly described in the next section. Special tasks are used in functional MRI (fMRI) to register brain activity during a particular action (e.g., reading). Visualization of our brains' work happens by imaging the change in blood flow related to energy use by brain cells (hemodynamic response), which is often called Blood-oxygen-level dependent contrast imaging (BOLD-response). This very precise mean to not only locate brain activity, but also to measure level of its reaction, is called timecourse, when measured over a period of time. The fundamental assumption of fMRI-research is that when an individual perform two tasks simultaneously, BOLD-response is added linearly, what means scaling (multiplying by a number) for each chosen process and their summation. Emerging redundant data, or noise, was successfully managed by gathering big amount of fMRI-records. That is why experiments with a lot of participants are much more valuable.

3.2 Using neurovisualization in linguistics

The relation between human brain and language system was always key question in understanding our nature. Neurolinguistics investigate questions, connected with brain reaction not on real world objects, but rather with organization and access to conceptual information, stored in a word. This is very important feature of neurolinguistic research, because neuroimaging data shows, that processes of word and object recognition are not at all equivalent: in the first case, activation of perceptual processes stimulates more and more abstract connections, while for the second, linguistic case, it is obvious that perceptual information can not be held in word form (colour characteristic of a flower is not contained in the word 'daisy').

A stimulus is the key concept in experimental linguistics and defines the very core of a research. Scientists use as stimulus an individual word (audible or written), a sentence (presented by words or as a whole), pictures (with sign or with task to create one), and continuous speech (e.g., audiobooks or parts of a dialogue). The listed above neurovisualization methods allow to solve problems, connected with different levels of language system - phonology, grammar, semantics, syntax, what defines design of an experiment and an neuroimaging technique.

The idea of experiment,described in Arshit Gupta (2015), is to associate MEG-data, showing brain activity in time resolution, with grammar information. And results of this research not only satisfuying: using Naive Bayes Classifier they achieve 88% accuracy in predicting part-of-speech tag for new stimulus and present coherent brain image. Undoubtedly, one of essential issue in connection with brain activity during natural language processing is perception of semantic information, i.e. speed of stimulus processing, main stages, localization. In order to investigate this issue, scientists use contrast between two groups of nouns - abstract(e.g. future) and concrete(e.g.,dog), because it allows inferences to be made as to how different kinds of information (i.e. linguistic versus perceptual/imaginal) contribute to word recognition. There are some theories on neuro basis for concrete-abstract opposition, and the most popular was a dual-coding theory, according to which processing of these two types of words were located in left hemisphere (for abstract words) and in right-hemispheric brain areas for concrete words (see Paivio (1990)). But recent researchers (Fiebach (2003)) claim the opposite: when individual were given lexical decision tasks and fMRI-data was gathered it was revealed that abstract words activated a subregion of the left inferior frontal gyrus more strongly than concrete words, but there were no difference in localization of brain activity during processing of both concrete and abstract words(not in hemisphere scale). However, their results revealed the level of BOLD-response was much more stronger when stimulus was an abstract word. Therefore it is essential to consider differences in time and intensity, not in localization (in cases of concrete-abstract opposition).

To some extend, main focus of almost every modern work in neurolinguistics is either on semantics or on syntax, but there is an interesting discoveries in comparison of this two language levels. It is revealed in Friederici (2000) that activation area depends not on semantic oppositions or syntax class, but on task itself - whether an individual should process semantic or syntax information. These results suggest that difference between two levels much more important than between two groups within one language level, that they are functionally distinct and involve different subparts of the neuronal network. Thus, word processing supports a domain-specific organization of a language.

4 A historical approach to linking neuroimaging data and computational linguistics models

As we see the growing rise of interest to cognitive sciences and neurosciences, it is no wonder that each year we see more and more interesting works related to these domains. One of these topics includes the investigation of how the human brain works with semantic information. Naturally, this domain attracted specialists in natural language processing, who apply various semantic models to crack open the black box and understand, what exactly is going on in the human brain when stimulus words are processed and what reaction they cause. In this section we provide a short overview of the works, where some of the aforementioned NLP models, that deal with semantics, are used to discover new patterns of semantic information processing.

4.1 Prediction-based models

The idea that the human brain has a complex structure for processing various information is not a new one. Many previous works, e.g.Shinkareva et al. (2008), introduced the idea that encoding of words in the human brain encompasses different brain parts and is partially shared across individuals, but the facts that these encodings are actually common for all the people and the existence of voxel clusters responding stronger to words with specific semantic categories, were not presented until 2008 and 2016, respectively. This part focuses on the results, acquired by word embedding model, one of the prediction-based models.

The first work that pioneered in this field was published by Tom Mitchel and colleagues Mitchell et al. (2008). The authors managed to establish a connection between a word's semantic category and the reaction of a specific group of voxels in response to this word. This was achieved by matching a word with an fMRI image of a participant's brain, taken when the stimulus word was presented. The approach introduced in this paper suggested that a voxel's activation was calculated as a weighted sum of semantic values of a given word multiplied by some scalar value, learned by the model introduced in this paper. A word was represented as a vector, with each i-th coordinate corresponding to the word's semantic values, namely, its co-occurrence with 25 sensory-motor verbs (e.g. see, smell, etc) in a large corpora. Next, the authors used these vector representations to create a predicting model, that produces the fMRI image for a brain activation in response to some word (any word, even those outside the training set).

It is very important to mention, that the experiment setting of the research is based on the following procedure: 60 nouns belonging to 12 semantic categories were presented to 9 participants in random order, 6 times each; then, a final fMRI response for each word was calculated as the mean of all responses to a given word. The fact is important as the further settings differ greatly.

The most striking result of this work is that the semantic center (brain parts responsible for processing semantic information) is more or less common across the participants in terms of its localization in a brain, and the reaction of each individual in response to the given stimuli doesn't differ significantly from the reaction of other participants. As the article presented a whole new approach to analysis of semantics and information processing in the human brain, the results triggered an interest of the scientific community to this domain, while the authors set the baseline for future works. Further experiments, e.g. Jelodar et al. (2010) and Devereaux et al. (2010), introduced new ways of improving the results acquired in Mitchell et al. (2008), such as using WordNet and grammatical relations as features for Mitchel's model.

The next milestone for the research in this field was set in the work Huth et al. (2016) published by Huth et al. The focus of this work was on creating an interpretable «semantic map» of a brain, revealing that each unit of the so-called semantic system is selective, that is, it responds stronger to words with specific characteristics, namely, the semantic ones.

The experimental setting comprised the following aspects: first, the subjects were listening to the story called «The Moth Radio Hour», while their BOLD responses were being recorded by fMRI. The first prominent fact about this research is that the previous studies were not conducted with «a natural» stimuli; only some separate words were used as stimulus words, while here the reactions were elicited by a coherent text . Next, a word embedding space for each word from the story was used to identify their semantic features, characterized by co-occurence with popular English words (985 linguistic items). These features were used for linear regression model to reveal how they modify BOLD responses. Finally, the words from the story were clustered; each category was given a semantic label for further work with «semantic atlas». Huth et al. introduced PrAGMATiC, an algorithm for constructing the «semantic atlas» of any person, which is actually more or less similar across the participants. The results were statistically significant and the data is valid. Thus, using word embedding model, the authors revealed a complex semantic structure of the human brain, shared by all the individuals.

One of the latest interesting results following Huth's and Mitchell's work can be found in Pereira et al. (2018), published by Pereira et al. The authors use their word embedding model to create a semantic space capable of generalizing to previously unseen data. Besides, Pereira et al. train their model on fMRI images and show, that their model is capable of decoding semantic vector from the image data. The most striking aspect of this research is that Pereira et al. do not confine the work to decoding separate words only, but instead work with higher level semantic representations, that is, with the meaning of phrases and sentences. The constructed decoder proved to be robust and could generalize to new concepts quite well.

Finally, it should be noted, that different results can be achieved by different word embedding models. A recent work published by Abnar et al. (seeAbnar et al. (2017)) focuses on comparing the results of different embedding models for the data taken from Mitchell et al. (2008). The motivation of the research in this domain can be explained by the fact that different models encode words in different ways, thus the importance of choosing the right model for future experiments arises. The authors' experiments included the use of Word2vec model and its variations, FastText, GloVe and LexVec. The results reveal, that Glove and dependency-based Word2Vec get the highest scores for the task set in Mitchell et al. (2008). The discussion on the effectiveness of current word embedding models also resulted in a prominent work by Fyshe et al. Fyshe et al. (2014). In this paper the authors point out the disadvantages of current word embedding models and claim, that vector semantic representations may become closer to ground truth, if word encoding data is combined with corpus-derived statistics. The hypothesis underlying the research is that vector models using statistics of word co-occurences are unable to fully grasp the semantic representations of a given word because of the following reasons: 1. Words with multiple meanings are collided in one vector; 2. Texts may contain spam; thus, brain activation data may enrich the information encoded in current models and improve their accuracy. Although this approach is somewhat debatable, since 1. one and the same stimulus word may elicit different reaction across the participants; 2. the brains of participants may differ in shape, once again, producing different data for the same stimulus, the authors reveal, that their approach provides good results and the semantic representation of words in their model is more accurate.

4.2 Language models

Statistical language models are another way of exploring word semantics and are very popular nowadays. Most of the modern statistical models are more or less related to the use of neural networks, and, since the latter provide state-of-the art results, we can conclude, that these models are quite effective.

Statistical language models are becoming extremely popular for the research of how words are encoded in our brain and what the reaction of brain to a word will be. For example, in Sudre et al. (2012) the participants were reading various nouns, and, as a result, each subject's MEG data was used to predict a word they were reading. Sudre et al. use this MEG data and statistical models to differentiate two different words in terms of semantic characteristics, which are believed to be encoded in neural activity, recorded by MEG. The use of MEG and EEG data is what constitutes a difference between the research related to word embeddings and the research involving statistical models, as in the latter it is claimed that language dynamics can be better represented by MEG and EEG, rather than fMRI.

Predictability of a given word(which is believed to be strongly related to a word's semantic features, see more details in the end of the chapter) is another popular application of statistical models. It was explored in Dambacher et al. (2006) by Dambacher et al. and in Frank et al. (2013) by Frank et al. In Dambacher et al. (2006) in order to get information on whether a word is predictable (predictability is described by the so-called cloze probabilities) was acquired by human annotators;

then an experiment measuring the activity of N400(a well-known predictor of word surprisal in terms of semantic properties a word) was run in order to find a correlation between N400 activity and cloze probabilities. The results proved the existence of such a correlation, but the conducted research was subjective, as the word probabilities were based on annotators' personal opinion, rather than more objective factors. This approach was later revised in Baroni et al. (2009), where each word's probability was estimated using statistical models, namely, Markov models, phrase-structure grammars and RNNs(Recurrent Neural Networks), trained on the British National Corpus. The results of the work show, that RNN outperforms all the previous models; Markov model's poor performance is explained by the fact, that it takes into account only a restricted context(from 2 to 4 words), while RNNs cover all the previous history of a given word. This data partially correlates with the data in Parviz et al. (2011), published by Parviz et al., who showed that starting from n=4 for Markov model the prediction results tend to improve, that is, the more history is taken into account, the better the results the model outputs.

The next interesting result involving statistical language models can be found in Wehbe et al. (2014), published by Wehbe et al. This research follows the experimental paradigm set in Sudre et al. (2012), but instead of working with separate words, the stimuli presented was more «naturalistic», namely it was a collection of texts. The authors draw a parallel between the way a neural network outputs a probability of a given word, taking the previous history into account, and the efforts the human brain uses to integrate a new word into the previous context, as the efforts are proportional to the word's probability given previous history (the aforementioned efforts are related to N400, see Frank et al. (2013) for more details on the N400 activity). Wehbe et al. use two statistical models: Recurrent Neural Network Language Model, introduced in Mikolov et al. (2011), and Neural Probabilistic Language Model (NPLM) to get the output probabilities of words from their dataset and at the same time run a MEG experiment on 3 subjects, who read the words from the very same dataset. While the first statistical model takes all the previous history into account, the second model is restricted to some small context. The results revealed, that RNNLM slightly outperforms NPLM in capturing text contexts, thus, showing better scores during testing.

Similar research related to reading was carried out by Mitchell et al. in Mitchell et al. (2010). The authors use eye-tracking data to explore reading process and gaze fixation in terms of semantic and syntactic constraints, namely, they try to predict reading time using these features. Other factors, such as word's length, word's frequency was also taken into account, but the main focus of the paper is still semantic and syntactic features. The assumption underlying the research is that the less surprising a word is, the lower the cognitive load for integrating it into context. One striking aspect of this paper is that the introduced LDA model works not with separate words, but with word sequences. The key result of the article is the following: the previous works on the same subject introduced a model, predicting text reading time using syntactic features only, while Mitchell and all showed, that semantic features are also important predictors, significantly improving the model's predictions.

Finally, in order to combine the existing approaches related to a word's predictability and its semantic features, a research was carried out by Frank et al. Frank and Willems (2017) to investigate the connection between predictability and semantics. Using statistical language model to compute a word's probability given the previous words and distributional semantic model to evaluate the semantic similarity between a given word and the previous words, Frank et al. reveal, that both predictability and semantic similarity are characterized by similar N400 effects, but dissimilar fMRI data across individuals.

It should be noted, that most of the previous works, mentioned here, were centered around working with one word and its additional information only. Thus, at a particular moment of time a new idea, namely, the prediction of semantically meaningful sentences using brain activations data, became popular. The first attempt to run such an experiment was made by Toneva et al. in Toneva et al. (2015). Following Wehbe's idea Wehbe et al. (2014), the authors use words' contexts derived from the hidden layer of recurrent neural network language models and integrate this information into their model to predict sentences using MRI data. Using three different models and custom metrics for the their performance evaluation, Toneva et al. conclude, that the best results are achieved by models, that take into account only context information, not the models with both context and semantic information.

4.3 Count-based models

In this part we focus on Latent Semantic Analysis (LSA). LSA is another method for exploring semantics that can be found in some works, related to semantic information encoding in human brain. While word embedding models use vectors and statistical models use probabilities to work with semantics, LSA is mostly related to matrix decompositions, namely, to Singular Value Decomposition.

For example, inDevereaux et al. (2010) Devereaux et al. used a model, introduced by Baroni et al. in Dambacher et al. (2006),Baroni et al. (2009), that is based on LSA and HAL. The initial goal of the paper was to derive fMRI image using word vectors. Following Mitchell's methodology from Mitchell et al. (2008) to some extent, the authors get interesting results: namely, while Mitchell et al. put emphasis on linking important semantic features with sensory-motor verbs, Devereaux et al. focus on intrinsic semantic features and produce their analysis of the derived components. Then the authors compare their model and Mitchell's model for fMRI image prediction; as a result, the first model outperformed the latter.

LSA was also used for exploring reading time. For example, in Pynte et al. (2008) Pynte et al. tried to use LSA for analyzing contextual influences. The authors want to evaluate the semantic constraints exerted on a given word by a prior word an a prior sentence using eye-tracker data. This can be done by measuring the similarity of 2 vectors: the vector of a given word and the vector for a word preceding it; the closer these two vectors are, the more similarity these words share and the less constraint is put on a given word. The same assumption is true for a prior sentence. As the meaning is connected with syntax, syntactic features are also taken into account. The results were in line with the previous works and reveal, that gaze duration and word predictability are connected, that is, the more a word is related to a previous one, the less time for gaze fixation is required.

The aforementioned experiments with N400 were also conducted using LSA. In 2011 Parviz et al. tried to evaluate the strength of N400 using the features derived by 3 different models, including LSA [15]. To be more precise, the authors tried to analyze, what feature contribute most to this or that psychological or neural response. Using MEG data, collected from reading experiment, they show, that LSA predictors are usually statistically more significant, than the predictors provided by other models, namely, the 4-gram statistical model and the Roark parser Roark (2001).

5 Future challenges

To conclude, our survey showed that studies trying to find an interplay between computational semantic models and empirical neurolinguistic data are able to obtain positive results, thus reporting that bridging between two these semantic paradigms is possible. Our work showed that distributional semantics is useful for psycholinguistic studies either psycholinguistics can be used in computational linguistics research through giving insights on how much semantics modeled in computer is close to the way of how humans process semantics in their brain.

From our perspective, it is not possible to say which option has the biggest potential for future studies and applications. We think that the biggest scientific breakthrough should happen on the intersection of fields of cognitive studies and computational modeling, and balanced development of both of them is equally important. Without computational methods we will not be able to construct a model of semantic (or even other parts of language) processing , to transfer to a machine the ability to process meaning of linguistic units. But, on the other hand, if we have a computational model, we should understand how our own brain works, and this is impossible without studying the structure of our cognition.

But another point that we also want to say is that it is important to understand that computational semantic models that were described in this study are not silver bullets. These models have their own limitations, and one should carefully treat the results obtained with experiments strongly relying on such models. For instance, in case of exploration of localization of semantic categories, we should understand that information about relations between these categories in a brain space is obtained through Word2Vec word embeddings model which very roughly treat the concept of semantics. And if we report certain conclusions using something so rough, should also our conclusions treated as rough? Of course, if we report results based on a use of somebody's neuroimaging data, we are just treating to the author of this data because we cannot verify the setup of experiments in which this instrument was obtained, so we cannot prove whether this data bad or good. But in the case with word embeddings we already know that this instrument is not really good. From this perspective particularly, the «neurolinguistics for computational linguistics» seems to be more reliable.

To mention certain future directions that we treat as ones with big potential, we could say about some of them.

Firstly, we think that studies on intersection of neurolinguistics and computational linguistics would be really actual for low-resource and poorly inverstigated languages, such as Chuvash. Processing of such languages is not only crucial for development of low ethnic communities. We also think that while we are starting to build a language-independent paradigm on a base of well-known (in the field of cognitive studies) languages, some of rare, underestimated languages could break some of our paradigms.

Secondly, we see that neuroimaging tools develop, becoming cheaper and better in quality. This could lead us to a situtation when fMRI scanner will be available to each university, and scans could be obtained with an interval of decimals of seconds. Possibly, this turn in neuroimaging can ruin the existing results — but also it can approve them.

Thirdly, trends in computational linguistics are moving to an a higher integration of linguistic knowledge to computational models. May be in future we will be able to see experiments with models that are not counting corpus-statistic, but also relying on certain manually constructed knowledge bases.

Intersection of computational semantics and neurolinguistics has a lot of bright future frontiers, but mentioning all of them goes beyond the scope of this study. Anyway, our survey showed that a research interests to this topic grows as well as researchers started to pose more and more complex questions. We can conclude that it will have a bright future, being one of the most crucial bricks in creation of artificial intelligence.

Acknowledgments

We want to thank our colleagues and academic supervisors, Alexey Artemov, Anastasiya Lopukhina and Svetlana Malyutina, for productive discussions on this work, critique and suggestions.

References

Abnar, S., Ahmed, R., Mijnheer, M., and Zuidema, W. (2017). Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. *arXiv preprint arXiv:1711.09285*.

Armando Freitas da Rocha, F. B. F. and Alfredo Pereira, J. (2015). Combining different tools for eeg analysis to study the distributed character of language processing. *Computational Intelligence and Neuroscience*.

Arshit Gupta, T. M. (2015). Learning to identify pos from brain image data. *Brain Image Analysis Group - Final Report*, pages 1–13.

Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2009). A corpus-based semantic model based on properties and types. *Cognitive Science*, 34:1—-333.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bloomfield, L. (1936). Language (new york: Henry holt, 1933). *Language or Ideas*, pages 89–95.

Brennan, Jonathan R.; Wagley, N. K. I. B. S. M. R. A. E. L.-O. R. (2016). Magnetoencephalography shows atypical sensitivity to linguistic sound sequences in autism spectrum disorder. *NeuroReport*, 27:982–986.

Collette, Majerus, V. d. L. e. (2001). Contribution of lexico-semantic processes to verbal short-term memory tasks: A pet activation study. *Memory*, 9(1):249–259.

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

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Cyma Van Petten, M. K. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition*, 18(4):380—-393.

Dambacher, M., Kliegl, R., Hofmann, M., and Jacobs, A. M. (2006). Frequency and predictability effect on event-related potentials during reading. *Brain Research*, 1084:89—-103.

Devereaux, B., Kelly, C., and Korhonen, A. (2010). Using fmri activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 70–78.

Fiebach, F. (2003). Processing concrete words: fmri evidence against a specific right-hemisphere involvement. *Neuropsychologia*, (3):62–70.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Frank, S., Otten, L. J., Galli, G., and Vigliocco, G. (2013). Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, pages 878–883.

Frank, S. L. and Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*.

Frege, G. (1892). Über sinn und bedeutung. zeitschrift für philosophie und philosophische kritik 100, 25–50. translated as. *On Sense and Meaning" in Frege (1984). Google Scholar*.

Friederici, Opitz, C. (2000). Segregating semantic and syntactic aspects of processing in human brain:an fmri investigation of different word types. *Neuropsychologia*, (10):698–705.

Fyshe, A., Talukdar, P. P., Murphy, B., and Mitchell, T. M. (2014). Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2014, page 489. NIH Public Access.

Grefenstette, G. (2012). *Explorations in automatic thesaurus discovery*, volume 278. Springer Science & Business Media.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Jelodar, A. B., Alizadeh, M., and Khadivi, S. (2010). Wordnet based features for predicting brain activity associated with meanings of nouns. In *roceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 18–26. ACL.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lesya Y. Ganushchak, Ingrid K. Christoffels, a. N. O. S. (2011). The use of electroencephalography in language production research: A review. *Frontiers in Psychology*, 2(208):249–259.

Mikolov, T., Kombrink, S., Deoras, A., Burget, L., and Cernocky, J. (2011). Rnnlm- recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop*, pages 196–201.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitchell, J., Lapata, M., Demberg, V., and Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206. ACL.

Mitchell, T. M., Shinkareva, S., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(May):1191–1195.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer.

Paivio, A. (1990). *Mental Representations: A dual coding approach*, volume 322. Oxford University Press.

Parviz, M., Johnson, M., Johnson, B., and Brock, J. (2011). Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 38–46.

Paulmann, Seifert, K. (2010). Orbito-frontal lesions cause impairment during late but not early emotional prosodic processing. *Social Neuroscience*, 5(1):59–75.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Botvinick, M., and Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *NATURE COMMUNICATIONS*, 9(March).

Pynte, J., New, B., and Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research*, 48(21):2172—-2183.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249—-276.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics*, 24(1):97–123.

Shinkareva, S., Mason, R. A., Malave, V. L., Mitchell, T., and Just, M. A. (2008). Experiential, distributional and dependency-basedword embeddings have complementary roles in decoding brain activity. *PloS One3*, 3(1)(Feb).

Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., and Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451–463.

Tarski, A. (1944). The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research*, 4(3):341–376.

Toneva, M., Singh, V., and Wang, H. (2015). Supervised mind reading: Uncovering text from neural data. *semanticscholar.org*, https://pdfs.semanticscholar.org/acbb/de42e1d77f2d09a2580a5e753091ccfcd406.pdf.

Wehbe, L., Vaswani, A., Knight, K., and Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243.