A Survey of Word Embeddings Evaluation Methods

Amir Bakarov

Institute for System Analysis of Russian Academy of Sciences (ISA RAS),
The National Research University Higher School of Economics,
Moscow, Russia
amirbakarov at gmail.com

Abstract. Word embeddings are real-valued word representations able to capture lexical semantics and trained on natural language corpora. Models proposing these representations have gained popularity in the recent years, but the issue of the most adequate evaluation method still remains open. This paper presents an extensive overview of the field of word embeddings evaluation, highlighting main problems and proposing a typology of approaches to evaluation, summarizing 16 intrinsic methods and 12 extrinsic methods. I describe both widely-used and experimental methods, systematize information about evaluation datasets and discuss some key challenges.

1 Introduction

Word embeddings, real-valued representations of words produced by distributional semantic models (DSMs), are one of the most popular tools in modern NLP, but their nature and limitations are still not well understood. One of the most important questions in the studies of distributional semantics is how to evaluate the quality of DSMs. There is still no consensus in the scientific community about which evaluation method should be used: NLP engineers who are more interested in dealing with downstream tasks (for instance, semantic role labeling) usually evaluate the performance of embeddings on such tasks, while computational linguists exploring the nature of semantics tend to investigate word embeddings through experimental methods from cognitive sciences.

The aim of this paper is to systematize and classify all existing approaches to the task of word embeddings evaluation, and by “all existing” I mean both widely-used (and at the same time widely-criticized) and less mainstream, experimental approaches. In this work, I suggest a hierarchical typology of word embeddings evaluation methods, highlighting the main problems and systematizing the existing datasets. I consider this paper to be the most extensive survey in the field of word embeddings evaluation as of now.

The paper is organized as follows. In Section 2 I describe the recent advances in word embeddings evaluation and briefly summarize the ideas proposed in the key works. Sections 3 and 4 are dedicated to the two primary classes of evaluation methods, extrinsic methods and intrinsic methods. In these sections I shortly explain how each of the proposed method in each of the sections works. Section 5 concludes the work, and there I propose my thoughts on some future challenges in the field of word embeddings evaluation.

2 Brief History

In this section I propose an overview of previous works dedicated to the task of word embeddings evaluation that seem to be important in this field. Notably, this section does not cover all the existing studies – more works would be considered in the following sections devoted directly to overview of evaluation methods.

The first work in which word embeddings evaluation was addressed (there were no word embeddings, though, so the similar concepts were called *distributional semantics*) was carried out in [Griffiths et al., 2007], even before the distributional semantics gained its main popularity. In 2010 a big survey of tasks that could be solved with the help of distributional semantics models was proposed (hence, all of these tasks could be considered as a measure of word embeddings performance) [Turney and Pantel, 2010]. A year after, in 2011, the first comparison of performance of various DSMs was published [McNamara, 2011]. In 2013, the popular *Word2Vec* tool was released, carrying out novel approaches to evaluation (like the word analogy task, $king - man + woman = queen$) [Mikolov et al., 2013a]. In 2014, [Baroni et al., 2014] proposed an extensive overview of approaches to word embeddings evaluation. Then, in 2015, [Schnabel et al., 2015] systematized the existing approaches into two major classes which are *extrinsic evaluation* and *intrinsic evaluation*. In 2016, the first workshop on word embeddings evaluation took place at the Annual Meeting of Association of Computational Linguistics (*RepEval 2016: The First Workshop on Evaluating Vector Space Representations for NLP*). This workshop provided a lot of interesting works and research proposals, with various aspects of word embeddings evaluation considered. Particularly, this workshop helped to highlight the most important problems in the field [Faruqui et al., 2016]:

1. **Obscureness of the notion of semantics.** Word embeddings are usually considered to be "good" if they reflect our understandings of semantics. But at the same time, we are not aware whether out understandings are absolutely correct. Moreover, it is unclear which type of relationships between words word embeddings should reflect, because there are many different types of relations between words (like semantic relatedness and semantic similarity, their definitions are notably also quite obscure). It is also not clear whether the model should be considered "bad" if it takes into account not the relationships between words that we used to define as a semantic relatedness or similarity.
2. **Lack of proper training data.** Most of the evaluation datasets are not divided into training and test sets. Hence, researchers training the word embeddings adjust them to the data trying to increase their quality. They are trying to capture not the actual relationships between words, but the relationships existing in the data.

3. **Absence of correlation between intrinsic and extrinsic methods.** Performance scores of word embeddings, when measured with two existing evaluation approaches (intrinsic and extrinsic), do not correlate between themselves. It is unclear what class of methods is more adequate.
4. **Lack of significance tests.** Statistical significance tests are sometimes not performed in the key experiments with new distributional models and evaluation methods. Thus, certain results of evaluation proposed in certain papers are not as correct as it is desirable.
5. **The hubness problem.** It is unclear how to deal with so-called *hubs* which are word vectors representing very frequent words. Such vectors are close to a disproportionately large number of other word vectors, hence, cosine distances between any two word vectors would probably be noised by the hubs, and the any evaluation in this case is biased.

Among other key papers presented at *RepEval-2016*, I would point at the overview of the problems of existing word embeddings evaluation datasets (subjectiveness of rating scales, lack of penalties for overestimating semantic similarity of two dissimilar words, etc.) [Avraham and Goldberg, 2016] and the overview of methods of intrinsic word embeddings evaluation (which, however, does not address all the methods presented in this survey) [Gladkova and Drozd, 2016]. Also, there were a lot of other significant works outside of *RepEval*, which raised important questions (the problem of the proper size of the dataset, uncertainty about which words to include in a representative dataset, etc.) [Jastrzebski et al., 2017]. In 2017, the second workshop was held (*RepEval 2017: The Second Workshop on Evaluating Vector Space Representations for NLP*), but I would not say that works presented there addressed such range of significant problems.

Thus, by efforts of a lot of researchers many questions in the task of word embeddings evaluation were raised. Some of these questions were already answered, but much more still remain open. I hope that this work could help to investigate and resolve the remaining problems.

3 Extrinsic evaluation

Methods of extrinsic evaluation are based on the ability of word embeddings to be used as the feature vectors of supervised machine learning algorithms (like Maximum Entropy Model) used in one of various downstream NLP task. The performance of the supervised model (being measured on a dataset for NLP task) functions as a measure of word embeddings quality. Some researches assume that word embeddings showing a good result on one task will show a good result on others, and the results of word embeddings on different tasks correlate, defining some kind of a global evaluation score for distributional semantics.

Word embeddings probably could be used in almost any NLP task, and certain researchers (while describing possible options of using word embeddings in other downstream tasks) do not mention the task of word embeddings evaluation. Nevertheless, by the definition of extrinsic evaluation given above, any downstream task could be considered as an evaluation method. In this section, I select and describe the following tasks used by other researchers for extrinsic evaluation either implicitly (without mentioning the problem of word embeddings evaluation) or explicitly (as in [Nayak et al., 2016], for instance). I do not try to mention all existing NLP tasks[1], but only the tasks in which I am aware the word embeddings were used.

1. **Noun Phrase Chunking**, to identify noun phrases and their boundaries within the sentence (i.e., to mark all the bigrams *noun + dependent word*). [Schnabel et al., 2015,Turian et al., 2010,Collobert et al., 2011]. I am aware of a dataset for this task prepared on a *CoNLL-2000* shared task [Tjong Kim Sang and Buchholz, 2000].
2. **Named Entity Recognition**, to identify types of named entities (names of organizations, people, brands, etc.) in the sentence and their boundaries [Turian et al., 2010,Collobert et al., 2011]. There are several datasets for this evaluation method, including datasets prepared for the *CoNLL-2002* and *CoNLL-2003* shared tasks [Tjong Kim Sang and De Meulder, 2003], and a dataset made for the *MUC-7* shared task [Chinchor and Marsh, 1998].
3. **Sentiment Analysis**, a particular case of a text classification problem, when a text fragment should be marked with a binary label reporting positive or negative polarity of the text sentiment [Schnabel et al., 2015,Tsvetkov et al., 2015]. There are certain user reviews datasets (for example, movie reviews [Maas et al., 2011]) that could be used for evaluation on this task.
4. **Shallow Syntax Parsing**, to decompose the sentence into phrase groups (not only noun phrases, but also verb phrases, adjective phrases, etc.) [Bansal et al., 2014,Köhn, 2016]. In some papers a similar task is called as **Parse Tree Level 0 Construction** [Collobert et al., 2011], in some papers the task is extent to the full dependency or constituent parsing task [Andreas and Klein, 2014]. Any datasets constructed for evaluation on such parsing task (like [Marcus et al., 1993]) could be used for evaluation.
5. **Semantic Role Labeling**, to identify thematic (semantic) roles of arguments for various predicates within the sentence [Collobert et al., 2011]. Some researchers formulate the task as a classification task which is to classify each sentence in a set by the thematic role of a specified word which occurs in an each sentence in the set [Ettinger et al., 2016a]. There are several datasets that could be used for the first type of the semantic role labeling task, like the *Proposition Bank* [Palmer et al., 2005].
6. **Negation scope**. This task also could be considered a text classification task. It is to identify whether a specified action in a sentence determined is a negation or not (such actions are contained in each sentence in a certain set) [Ettinger et al., 2016a]. However, there are no existing datasets that could be used for evaluation on this task.
7. **Part-of-Speech Tagging**, to identify part of speech of each word in the sentence [Collobert et al., 2011]. Part-of-speech datasets could be

[1] like at `https://aclweb.org/aclwiki/State_of_the_art`

considered suitable datasets for this task, like the *Stanford Treebank* [Toutanova et al., 2003].

8. **Text Classification**, which is in general to mark a text fragment with a label depending on its content: for example, categorizing sport news based on the type of sport activity they are about (football or basketball) [Tsvetkov et al., 2015]; there exist various datasets with text classified by their semantics, for instance, *20 Newsgroups*[2].
9. **Metaphor Detection** is another classification task which is to identify whether the specified phrase (like *adjective-noun* or *subject-verb-object*) is metaphorical or literal [Tsvetkov et al., 2014,Tsvetkov et al., 2015]. I am aware of two datasets for this task, which are *Trope Finder Dataset*[3] and the dataset proposed by Yulia Tsvetkov in a study dedicated to metaphor detection [Tsvetkov et al., 2014].
10. **Paraphrase Detection** (also called **duplicate detection**, **paraphrase identification**, **record linkage**, **approximate string matching**, **text-to-text similarity detection**) is to determine whether two text fragments are paraphrases of each other (however, the notion of paraphrase is not so clear, and different researchers define this relation in different ways). The pair of text fragments could be labeled by a score reporting degree of text similarity or by a binary mark reporting the existence of a similarity [Baumel et al., 2016,Bakarov and Gureenkova, 2017]. There are several datasets for this task, including the *Microsoft Research Paraphrase Corpus* [Dolan and Brockett, 2005] and the *Quora Question Pairs Dataset*[4].
11. **Textual Entailment Detection** (also called **natural language inference task**). The task is in a some way similar to the previously mentioned paraphrase detection task since it is also to label a pair of two text fragments. But this task, however, is to identify whether one of these fragments is a continuation of another (so the relationship is not bi-directional) [Baumel et al., 2016]. Among datasets designed specifically for evaluation on textual entailment detection task one can mention the *Sentences Involving Compositional Knowledge* dataset [Marelli et al., 2014] and the *Stanford Natural Language Inference* corpus [Bowman et al., 2015]. Notably, certain researchers consider a very similar task suitable for word embeddings evaluation which is to pick one of two possible one-sentence continuations for a short story of several sentences (**Story Cloze Task**). There is a special dataset for this type of tasks named *Story Cloze Dataset* [Mostafazadeh et al., 2016].
12. **Input for artificial neural networks**. I put in a separate category the methods in which word embeddings are used as initial weights in the input layers for various types of artificial neural networks that are later employed to resolve downstream tasks like machine translation, morphological analysis and language modeling. Authors of this idea did not consider the problem of word embeddings evaluation, but they compared different types of DSMs,

[2] http://qwone.com/~jason/20Newsgroups

[3] http://www.cs.sfu.ca/~anoop/students/jbirke

[4] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

so I believe such a method could also be used as a framework for evaluation [Kocmi and Bojar, 2017].

However, I argue that there are significant issues in this group of methods. Of course, extrinsic evaluation has certain advantages, and in the cases where word embeddings are supposed to be used *only* to resolve a specific downstream task, evaluation of performance of supervised model on this task will give the most adequate score of word embeddings performance. But extrinsic evaluation fails if the embeddings that one wants to evaluate are trained to serve in a wide range of different tasks since word embeddings performance scores in various downstream tasks *do not correlate* between themselves (as it was hypothesized earlier) [Schnabel et al., 2015]. However, this fact is not surprising, since different downstream tasks differ very much, and they use completely different features of word embeddings (like, embeddings which are used in a POS-tagging task should consider words with the same part of speech to be similar independently of their semantics). Hence, no global evaluation score for word embeddings could be obtained through extrinsic evaluation methods.

Additionally, some researchers also highlight (among other shortcomings of extrinsic methods) high complexity of creating gold standard downstream tasks datasets (and, after all, in such datasets the possible subjectivity of human assessors is always an issue).

I argue that the ideas of extrinsic evaluation could be useful if one wants to highlight advantages of a certain DSM (showing its performance on a specific downstream task), but, due to the lack of performance correlation on different downstream tasks, such techniques can not be used as an absolute metric of word embeddings quality.

4 Intrinsic evaluation

Methods of intrinsic evaluation are experiments in which word embeddings are compared with human judgments on words relations. Manually created sets of words are often used to get human assessments, and then these assessments are compared with word embeddings (this method of collecting the judgments is called **absolute intrinsic evaluation**). The collection of assessments could be conducted either in the laboratory on a limited set of examinees (**judgments collected in-house**) or on crowd-sourcing Web platforms like *Mechanical Turk*, attracting an unlimited number of participants (**judgments collected through crowd-sourcing**) [Liza and Grzes, 2016].

Sometimes the assessors are asked to evaluate the quality of word embeddings directly, for instance, when different models produce different judgments on word relations, and the task of an assessor is to tell which model works better (such a method is called **comparative intrinsic evaluation**) [Schnabel et al., 2015]. Comparative intrinsic estimation allows not to estimate the absolute quality of embeddings, but to find the most adequate embeddings in a given set.

Absolute intrinsic evaluation uses *in vivo* experiments to obtain human judgments from assessors. The process of collecting the same data from word embeddings could be called *in vitro* experiment. The evaluation process is that two datasets obtained through these experiments are compared, and an aggregated estimate (for example, Spearman correlation coefficient) is calculated. Such an estimate could be used as an absolute measure of the quality of embeddings since it reports the similarity of lexical semantics inferred by embeddings to the lexical semantics determined by humans.

Most of methods of absolute intrinsic evaluation are designed to collect assessments which are results of conscious processes in a human brain (in other words, assessors have time to think about their answers). Hence, there is a probability that such answers are biased by certain subjective factors (for example, due to the absence of a clear definition of meaning every person interprets words relations in her own way, introducing the variability to the estimates). And it is not clear if the conscious assessments are really able to report the structure of semantics in a natural language.

Some researchers argue that such structure lies somewhere in the so-called subconscious level of cognition [Kutas and Federmeier, 2011]. If one could collect assessments directly from this level, then the evaluation probably will be far less biased. The attempts of collecting such data are becoming common in the field of word embeddings evaluation, and the evaluation becomes more interdisciplinary. Novel approaches to evaluation are based on using various neuroimaging methods which were previously used only in a field of psycholinguistics.

To this end, I would like to propose a new extensive typology of intrinsic word embeddings evaluation methods. In this typology I divide all existing methods into methods of **intrinsic conscious evaluation** and methods of **intrinsic subconscious evaluation**. This typology is inspired by the classification of data collection methods in psycholinguistic research which proposes *off-line methods*, in which examinees have time to think about decisions, and *on-line methods*, in which reflective responses are being collected.

However, I am also aware of other types of intrinsic evaluation methods that do not fall into these categories. Such methods are based not on a comparison with results of *in vivo experiments*, but on a comparison with knowledge bases, semantic networks and thesauri manually constructed by professional linguists and ontology engineers. I put these methods into another class, methods of **intrinsic thesaurus-based evaluation**.

Additionally, I argue that there is one more possible category of evaluation methods which is based neither on *in vivo* experiments nor on knowledge bases. This category proposes methods that are based on a comparison with data underlying in a language itself. Such data could be found, for instance, in graphematic representations of words, in speech sound signals or in frequency of occurrence of a pair of words in a corpus. I call them methods of **intrinsic language-driven evaluation**.

It is not obvious why the last three categories of the methods (proposed by myself) can be called evaluation methods at all, and I would like to explain this.

In other works researchers tend to use the notion of *exploration of word embeddings* for such experiments since it is unclear whether the data they use as the gold standard really contains information about lexical semantics. But is the existence of lexical semantics information contained in consciously assessed word similarity datasets truly clear? If one calls one type of experiments *exploration*, and calls another type *evaluation*, what is that threshold of evaluation method representativeness that would allow us to fit some methods into the evaluation category, and others into the exploration category? And then what is the definition of an evaluation method in the field of distributional semantics? After all, if the notion of word meaning could not be even defined properly, how the notion of its modeling evaluation could be defined?

My message is that I suggest to call a method of word embeddings evaluation any technique of finding embeddings correlation with ***any** data that hypothetically could carry information about lexical semantics*. Of course, the representativeness of such method will be questioned depending on the degree of plausibility of the hypothetical amount of necessary information in this data. But now we are not able to correctly evaluate this amount, and I argue that one should not say "this is not evaluation" just because of the lack of our capabilities to properly estimate the actual amount of semantic information contained in our data.

To this end, I have highlighted four classes of absolute intrinsic evaluation:

1. Methods of conscious evaluation;
2. Methods of subconscious evaluation;
3. Thesaurus-based methods;
4. Language-driven methods.

Now I am going to discuss more properly the design of each evaluation method proposed by other researchers which I fit into one of these classes.

4.1 Conscious intrinsic evaluation

4.1.1. Word semantic similarity method is based on an idea that the distances between words in an embedding space could be evaluated through the human heuristic judgments on the actual semantic distances between these words (e.g., the distance between *cup* and *mug* defined in an continuous interval $0, 1$ would be 0.8 since these words are synonymous, but not really the same thing). The assessor is given a set of pairs of words and asked to assess the degree of similarity for each pair. The distances between these pairs are also collected in a word embeddings space, and the two obtained distances sets are compared. The more similar they are, the better are embeddings [Baroni et al., 2014].

This method is one of the most popular methods for evaluation nowadays. Its roots go back to 1965 when the first experiments on human judgments on word semantic similarity were conducted to test the distributional hypothesis [Rubenstein and Goodenough, 1965] (in 1978 a similar work was carried out in [Osgood et al., 1978]). However, despite the strong psycholinguistic background of this method, it is one of the most frequently criticized in the community.

Critique to the method of word similarity has started no later than in 1999 (even before the notion of word embeddings appeared) [Friedman and Amoo, 1999]. The bulk of criticism addresses the subjective factor of such judgments which I mentioned earlier [Faruqui et al., 2016,Batchkarov et al., 2016]. By the claims of several researchers, there are more than 50 potential linguistic, psychological and social factors, which could introduce bias in the assessments [Gladkova and Drozd, 2016]. In some papers the problem of connotative words was also addressed: while denotative words (neutral common notions) do not address any assessors' associations, connotative words tend to cause subjectivity based on cultural or personal criteria [Liza and Grzes, 2016]). Other researchers also criticize the ambiguity of the task, since different experiments tend to propose different definitions of semantic similarity: some researchers define it as co-hyponymy (like the relation between the words *machine* and *bicycle*) [Turney and Pantel, 2010], while others define it as synonymy (like in a word pair *mug* and *cup*) [Hill et al., 2016].

It was also argued that the notion of semantic similarity inherits not only semantic connections of words, but also some morphological and graphematic features of word representations [Kiela et al., 2015]. Among other criticized features of word semantic similarity method are the lack of correlation between these human assessments and the performance of word embeddings on extrinsic methods [Chiu et al., 2016,Tsvetkov et al., 2015] (other researchers, however, explain this by the fact that such assessments are not sufficiently representative [Camacho-Collados and Navigli, 2016]), low inter-rater agreement between annotators [Hill et al., 2016], the factor of assessors getting tired when annotating a large number of pairs [Bruni et al., 2014], poor ability of numerical labels to fully describe all types of relations between words (it is suggested that it will be better to describe the degree of word similarity in a natural language [Milajevs and Griffiths, 2016]), and the mis-conduction of thematic roles relations [Erk, 2016].

Systematizing the results accumulated by other researchers, I propose all the datasets designed for evaluation in the task of word semantic similarity ranked by their size. Notably, different datasets use different notions of lexical semantic similarity, so the same embeddings could have different results on different datasets:

1. **SimVerb-3500**, 3 500 pairs of verbs assessed by semantic similarity (that means that pairs that are related but not similar have a fairly low rating) with a scale from 0 to 4 [Gerz et al., 2016].
2. **MEN** (acronym for Marco, Elia and Nam), 3 000 pairs assessed by semantic relatedness with a discrete scale from 0 to 50 [Bruni et al., 2014].
3. **RW** (acronym for Rare Word), 2 034 pairs of words with low occurrences (rare words) assessed by semantic similarity with a scale from 0 to 10 [Luong et al., 2013].
4. **SimLex-999**, 999 pairs assessed with a strong respect to semantic similarity with a scale from 0 to 10 [Hill et al., 2016].

5. **SemEval-2017**, 500 pairs assessed by semantic similarity with a scale from 0 to 4 prepared for the *SemEval-2017 Task 2* (*Multilingual and Cross-lingual Semantic Word Similarity*) [Camacho-Collados et al., 2017]. Notably, dataset contains not only words, but also collocations (*e.g. climate change*).
6. **MTurk-771** (acronym for Mechanical Turk), 771 pairs assessed by semantic relatedness with a scale from 0 to 5 [Halawi et al., 2012].
7. **WordSim-353**, 353 pairs assessed by semantic similarity (however, some researchers find the instructions for assessors ambiguous with respect to similarity and association) with a scale from 0 to 10 [Finkelstein et al., 2001].
8. **MTurk-287**, 287 pairs assessed by semantic relatedness with a scale from 0 to 5 [Radinsky et al., 2011].
9. **WordSim-353-REL**, 252 pairs, a subset of WordSim-353 containing no pairs of similar concepts [Agirre et al., 2009].
10. **WordSim-353-SIM**, 203 pairs, a subset of WordSim-353 containing similar or unassociated (to mark all pairs that receive a low rating as unassociated) pairs [Agirre et al., 2009].
11. **Verb-143**, 143 pairs of verbs assessed by semantic similarity with a scale from 0 to 4 [Baker et al., 2014].
12. **YP-130** (acronym for Yang and Powers), 130 pairs of verbs assessed by semantic similarity with a scale from 0 to 4 [Yang and Powers, 2006].
13. **RG-65** (acronym for Rubenstein and Goodenough), 65 pairs assessed by semantic similarity with a scale from 0 to 4 [Rubenstein and Goodenough, 1965].
14. **MC-30** (acronym for Miller and Charles), 30 pairs, a subset of RG-65 which contains 10 pairs with high similarity, 10 with middle similarity and 10 with low similarity [Miller and Charles, 1991]. Also, there is a subset of MC-30 called **MC-28** which excludes 2 pairs not represented in WordNet [Resnik, 1995].

4.1.2. Word analogy method (in certain works also called *analogical reasoning*, *linguistic regularities* and *word semantic coherence*) is the second most popular method of word embeddings evaluation. It is based on the idea that arithmetic operations in a word vector space could be predicted by humans: given a set of three words, a, $a*$ and b, the task is to identify such word $b*$ that the relation b:$b*$ is the same as the relation a:$a*$ [Turian et al., 2010,Pereira et al., 2016,Baroni et al., 2014]. For instance, one has words $a = Paris$, $b = France$, $c = Moscow$. Then the target word would be *Russia* since the relation $a : b$ is *capital* : *country*, so one needs need to find the capital of which country is *Moscow*.

The main criticism to this method addresses the lack of a precise evaluation metric. If in the word semantic similarity task the cosine distance between word vectors was intuitively adequate, then in this task the adequateness of such metric for relationship transfer is questioned. I am aware of three metrics used in the word analogy task:

– *3CosAdd* (and a similar metric *3CosMul*) proposed in the original *Word2Vec* paper is based on arithmetic operations in vector space (addition and multiplication of cosine distances) [Mikolov et al., 2013b].
– *PairDir* modifies *3CosAdd*, taking into account the direction of the resulting vectors in these operations [Levy and Goldberg, 2014].
– *Analogy Space Evaluation* metric compares the distances between words directly without finding the nearest neighbors [Che et al., 2017].

I also provide a list of datasets which could be used for evaluation on this task. As [Gladkova et al., 2016] notes, datasets designed for *semantic relation extraction task* could also be used to compile a word analogy set. In this case, it also worth looking at the *Lexical Relation* set which is a compilation of different semantic relation datasets including *BLESS* [Baroni and Lenci, 2011] (12 458 word pairs with a relation comprising 15 relation types) [Vylomova et al., 2015] and the *Semantic Neighbors* set (14 682 word pairs with a relation comprising 2 relation types, meaningful and random) [Panchenko et al., 2013].

1. **WordRep**, 118 292 623 analogy questions (4-word tuples) divided into 26 semantic classes, a superset of *Google Analogy* with additional data from WordNet [Gao et al., 2014].
2. **BATS** (acronym for Bigger Analogy Test Set), 99 200 questions divided into 4 classes (*inflectional morphology*, *derivational morphology*, *lexicographic semantics* and *encyclopedic semantics*) and 10 smaller subclasses. [Gladkova et al., 2016].
3. **Google Analogy** (also called Semantic-Syntactic Word Relationship Dataset), 19 544 questions divided into 2 classes (*morphological relations* and *semantic relations*) and 10 smaller subclasses (8 869 semantic questions and 10 675 morphological questions) [Mikolov et al., 2013a].
4. **SemEval-2012**, 10 014 questions divided into 10 semantic classes and 79 subclasses prepared for the *SemEval-2017 Task 2* (*Measuring Degrees of Relational Similarity*) [Jurgens et al., 2012].
5. **MSR** (acronym for Microsoft Research Syntactic Analogies), 8 000 questions divided into 16 morphological classes [Mikolov et al., 2013b].
6. **SAT** (acronym for Scholastic Aptitude Test), 5 610 questions divided into 374 semantic classes [Turney et al., 2003].
7. **JAIR** (acronym for Journal of Artificial Intelligence Research), 430 questions divided into 20 semantic classes. Notably, dataset contains not only words but collocations (like *solar system*) [Turney, 2008].

4.1.3. Thematic fit method evaluates the ability of a model to separate different thematic roles of arguments of a predicate (also called *selectional preference* [Baroni et al., 2014]). The idea is to find how well word embeddings could find most semantically similar noun for a certain verb that is used in a certain role. For humans, a certain verb could cause a person to expect that a certain role must be filled with a certain noun (e.g., for the argument *to cut* the most expected argument in the *object* role is *pie*) [Sayeed et al., 2016]. Experiments propose assessments of adequacy score of the tuple $verb, noun, thematicrole$ (for example, *people eat* is more common phrase than *eat people*, so the pair *people* and *eat* would have the higher score) [Vandekerckhove et al., 2009].

Some researchers consider another variation of this method, proposing the task of assessing a pair of words n (noun) and v (verb) by the most frequent role in which n used with v (e.g., pair $people, eat$ would be classified as the *subject* since it is more common to use *people* as a subject with that verb) [Baroni and Lenci, 2010].

In my opinion, the main problem of this method lies in two of its features. First, it needs a corpus annotated with thematic roles. Second, it is unclear which method of obtaining an embedding for a thematic role to distinguish different roles of the argument is the most adequate. Some researchers propose a method of vectorization of “slots” for certain thematic roles, which are obtained by averaging several most frequent nouns encountered in a given role – but applicability of such method is not obvious [Baroni and Lenci, 2010].

The following datasets could be used for evaluation with the thematic fit task:

1. **MSTNN** (abbreviation mentioned in [Sayeed et al., 2016]), 1 444 *verb-object-subject* pairs [McRae et al., 1997].
2. **GDS** (acronym for Greenberg, Sayeed and Danberg), 720 *verb-object* pairs. The dataset is additionally divided into a subsample containing only polysemous verbs (*GDS-poly*) and a subsample containing monosemous verbs (*GDS-mono*) [Greenberg et al., 2015].
3. **F-Inst & F-Loc** (acronym for Ferretti-Instrument and Ferretti-Location), 522 verbs pairs which are split to a subset of 248 verbs with associated *instruments* (*F-Inst*) and a subset of 274 verbs with associated *locations* (*F-Loc*) [Ferretti et al., 2001].
4. **P07** (acronym for Pado), 414 *verb-object-subject* pairs [Padó, 2007].
5. **UP** (acronym for Ulrike and Pado), 211 *verb-noun* pairs, the set of roles is unlimited [Padó and Lapata, 2007].
6. **MT98** (acronym for McRae and Tanenhaus), a subset of 200 verbs from *MSTNN* where each verb has two nouns, one is a good subject, but a bad object, and one which is a good object, but a bad subject [McRae et al., 1998].

4.1.4. Concept categorization method (also called *word clustering*) evaluates a word embeddings space to be clustered. Given a set of words, the task is to split it into subsets of words belonging to different categories (for example, for words *dog*, *elephant*, *robin*, *crow* the first two make one cluster which is *mammals* and the last two form another second cluster which is *birds*; the cluster name is not necessary to be formulated) [Baroni et al., 2014]. The amount of clusters should be defined. Possible critique of such method could address the question of either choosing the most appropriate clustering algorithm or choosing the most adequate metric for evaluating clustering quality.

Below I enumerate datasets which could be used as a gold standard for the task of word categorization:

1. **BM** (acronym for Battig and Montague), 5 321 words divided into 56 categories [Baroni et al., 2010].
2. **AP** (acronym for Almuhareb and Poesio), 402 words divided into 21 categories [Almuhareb, 2006].
3. **BLESS** (acronym for Baroni and Lenci Evaluation of Semantic Spaces), 200 words divided into 27 semantics classes [Baroni and Lenci, 2011]. Despite the fact that BLESS was designed for another type for evaluation, it is also possible to use this dataset in a word categorization task, as in [Jastrzebski et al., 2017].
4. **ESSLLI-2008** (acronym for the European Summer School in Logic, Language and Information), 45 words divided into 9 semantic classes (or 5 in less detailed categorization); the dataset was used in a shared task on a *Lexical Semantics Workshop on ESSLI-2008* [Baroni et al., 2008].

4.1.5. Synonym detection method, such as the *word semantic similarity method*, tries to evaluate the ability of word embeddings to form a vector space with predictable distances between words, but it does not propose an absolute degree of similarity: it is based on the idea that word similarity could be measured through finding the most similar word relative to a set of other words. Given a word a and a set $K = b_1, b_2, b_3$, the task is to find b_i which is the most synonymous (semantically similar in terms of the word semantic similarity task) to a [Baroni et al., 2014]: for example, for the target *levied* one must choose between *imposed* (correct), *believed*, *requested* and *correlated*. The task of a DSM is to find the word vector with the smallest distance to the vector of the specified word.

Taking into account all the criticism of the word semantic similarity method, moving from the absolute measure to the relative measure could probably exclude a lot of problems of this task (score bias, lack of assessments interpretability, etc). On the other hand, the creation of a dataset for evaluation in this task is more complicated and raises certain new questions (for example, how to properly choose the words to form the set K).

Datasets that could be used for evaluation on this task are listed below. They are presented in a form of 5-word tuples in which one word is a target word, and 4 words are potential synonyms where the only one is a correct answer.

1. **RDWP** (acronym for Reader's Digest Word Power Game), 300 synonym questions (5-word tuples) [Jarmasz and Szpakowicz, 2004].
2. **TOEFL** (acronym for Test of English as a Foreign Language), 80 questions [Landauer and Dumais, 1997].
3. **ESL** (acronym for English as a Second Language), 50 questions [Turney, 2001].

4.1.6. Outlier word detection method evaluates the same feature of word embeddings as the word categorization method (it also proposes clustering), but the task is not to divide a set of words into certain amount of clusters, but to identify a semantically anomalous word in an already formed cluster (for example, for a set $\{orange, banana, lemon, book, orange\}$ which are mostly fruits, the word *book* is the outlier since it is not a fruit) [Camacho-Collados and Navigli, 2016].

Some researchers propose a very similar method called *evaluation of coherence in semantic space*. The idea of this method is, given a set of three words – word a, the two words a_1 and a_2 which are the closest to a in an embedding space are found, – a word b is chosen randomly from the model's dictionary (this word probably would not be so semantically similar to a), and the task of a human assessor is to correctly identify b (the outlier) in the set a, a_1, a_2, b [Schnabel et al., 2015]. The more words are identified correctly, the better is the model.

There are two publicly available datasets for evaluation on this task:

1. **8-8-8 Dataset**, 8 clusters, each is represented by a set of 8 words with 8 outliers [Camacho-Collados and Navigli, 2016].
2. **WordSim-500**, 500 clusters, each is represented by a set of 8 words with 5 to 7 outliers [Blair et al., 2016].

4.2 Subconscious intrinsic evaluation

4.2.1. Semantic priming evaluation method is based on the same-name psycholinguistic behavioral experiment. It hypothesizes that a human reads a word faster if it is preceded by a semantically related word. The idea of an experiment is to measure the time of reading a specified word a (called the *target word*) in case it occurs after a word b_1 and in case it occurs after a word b_2. If the reading time of the word b_1 is lower than the reading time of the word b_2, than the word b_1 is claimed to be semantically related to a (b_1 is called *prime*, or *prime word*, or *stimulus word*) [Ettinger and Linzen, 2016,Auguste et al., 2017]. The time of reading could be obtained with the help of eye-tracking or safe-paced reading [Mandera et al., 2017,Lapesa and Evert, 2013], [Jones et al., 2006,Herdağdelen et al., 2009,McDonald and Brew, 2004].

I am aware of only one dataset that could be used for evaluation on the semantic priming task. It is the *Semantic Priming Project*, containing 6 337 pairs of words. The data is collected from 768 subjects for 1 661 target words. Every word pair presented in four versions: first, depending of the time interval on the demonstration of the target and non-target words which is 70 and 200ms (this interval is called *stimulus onset asynchronies, SOA*), and, second, depending on the task for the priming, naming task or lexical decision task [Hutchison et al., 2013].

4.2.2. Neural activation patterns When a person reads words, their meanings are hypothetically reflected in some patterns in her brain. Thus, such patterns could be used as input data for word embeddings. However, the consistency of such brain data is questioned since the neural activation patterns do not correlate in a large number of subjects (because the size and structure of the brain differ in different subjects). Another issue is that it is not clear to what extent these patterns have to do with lexical semantics, and with other linguistic data contained in words, like the number of characters, stress location, the number of syllables, etc. In the end, even the state-of-the-art techniques of neuroimaging are expensive, and it is unlikely that in the near future brain activity data can completely replace statistical corpus data. Nevertheless, some researchers propose that explorations of neural activation patterns could help to shed light on the nature of semantics, and therefore, to consistently evaluate the existing methods of word embeddings evaluation.

- **Functional Magnetic Resonance Imaging (fMRI)** evaluation method is based on using as a gold standard the data of the same-name neuroimaging experiment which measures changes associated with blood flow in certain parts of the brain by fixating regions of the blood flow at certain time intervals (once a second, for instance). The idea is that the blood flow and the neuronal activation patterns correlate, so one could identify parts of brain which are activated. In the field of neurolinguistics, reading or listening the text is usually considered to be a stimulus for this activity. The obtained data is presented as a set of voxels reporting the level of neuronal activity in different small parts of brain. It is not clear how to obtain data on reading single words, since the minimum time interval on fixating blood flow is about 1 second; some researchers try to train a regression model to compute the average brain activation vectors for each word or to use aggregate statistics to obtain vector representations of fMRI data using it is as a gold standard [Huth et al., 2016,Søgaard, 2016,Abnar et al., 2017]. One could try to use *StudyForrest* [Hanke et al., 2014] dataset which offers data on listening to the audio track of the “Forrest Gump” movie in German, or the *Word Processing* dataset which contains readings for various natural language words on English [Duncan et al., 2009].
- **Electroencephalography (EEG)** evaluation method is another method based on using neuroimaging data as a gold standard. Electroencephalography records the electrical activity of the brain, and the idea is that the amplitude of the impulses in the brain that occur on words (such response is called N400, it is an early response elicited by every word of a sentence) stores information about lexical semantics since the interpretation of the response is usually generalized by the hypothesis that the worse the word fits to the context (which could be both sentence context and word context), the higher is the amplitude of the signal. The amplitude differences of a tuple of words is able be simulated through the average cosine distances of word embeddings, so it is hypothetically could be used as a gold standard data for evaluation [Parviz et al., 2011,Ettinger et al., 2016b]. However, to this moment I am not aware of any publicly available EEG dataset that could be used for evaluation on this task.

4.2.3. Eye movement data evaluation method is based on using as a gold standard the data of human eye movement obtained. Such data could be obtained through instrument called *eye-tracker* tracks the movement of a pupil and a time of fixation on certain words while a person reads text from the computer screen, and such data hypothetically could carry some information about lexical semantics. The eye-tracker assigns to each word a set of features reporting characteristics of its reading: how many milliseconds the gaze was fixated on this word, how many times the gaze returned to it, etc. It is assumed that such feature sets could be compared with word embedding vectors by converting them to the vectors of aggregate statistics, and hypothetically the correlation between such vectors and word embeddings (for instance, on predicting k nearest neighbors to a certain word) could report the quality of a DSM [Søgaard, 2016].

I am aware only of two publicly available English eye movement datasets that one could use in their experiments. The first is the the **Provo Corpus** [Luke and Christianson, 2017] which consists of data of reading 55 paragraphs from 84 native speakers. This dataset could be converted in a list of 1 185 words each of which is associated with a set of 26 eye movement features.

The second dataset is the **Ghent Eye-Tracking Corpus (GECO)** [Cop et al., 2017] containing data of reading 5 000 sentences from monolingual and bilingual English speakers (33 participants overall). After converting one could obtain a dataset of 987 words, each associated with 48 features.

4.3 Thesaurus evaluation

4.3.1. Thesaurus vectors evaluation method (called **QVEC** in the original paper [Tsvetkov et al., 2015]) is based on the idea that word embeddings can be evaluated with the vectors of the inverted index of a collection of documents (*"thesaurus vectors"*) in a which each is responsible for a certain category of human knowledge, like super-senses in WordNet (e.g. *food*, *animal*, etc). The dimensionality of the thesaurus vectors is equal to the size of collection, and each component reports the number of occurrences of the word in a certain document; if the collection is too big, it is possible to use some kind of dimensionality reduction and map one component of an embeddings vector to multiple components of thesaurus vectors (or vice versa if the collection is too small). So, the gold standard is represented by the thesaurus vectors.

I believe that any set of documents which claims to contain comprehensive set of knowledge categories, can be used for evaluation (not only the conceptual thesaurus of WordNet super-senses). The most extensive one is *Wikipedia*, which is used for document vectorization with a similar thesaurus vector-based technique called *Explicit Semantic Analysis* [Gabrilovich and Markovitch, 2007]. It is usually applied in cross-language information retrieval.

4.3.2. Dictionary definition graph evaluation method is based on the idea that co-occurrences of words in dictionary definitions could carry information about their relationships [Acs and Kornai, 2016]. A digraph of a set of dictionaries in which the nodes are words could be constructed, and the value of the edge connecting the word a to the word b is equal to the number of occurrences of the word b in all definitions of the word a. This graph could be transformed into matrix, and for each word a *dictionary vector* could be obtained. Such vectors could be used as a gold standard in word embeddings evaluation.

Another variation of this graph exists, when the weights on the edges are not the frequencies of occurrences, but the numbers of using b as a head in the dependency syntax tree (such an idea can help to identify similarities based on phrases like *a cat is an animal*).

So, any dictionary can be used as a gold standard dataset for this task.

4.3.3. Cross-match test evaluation method is based on the same-name technique of finding similarity between two high-dimensional sets used to compare blood samples in medicine based on determining whether these two sets are sampled from the same distribution. Being applied to word embeddings evaluation, this method claimed to measure statistical significance of a model. Using it on two sets of word vectors trained on the same corpus, one could compute the correlation between these sets using the cross-match test. If the correlation is low, then the two compared models probably use different features of the corpus, so it is probably a good result [Gurnani, 2017].

4.3.4. Semantic difference evaluation method is based on using the characterizing words of distinctive features (called *attributes*). Each word in a pair is associated with a certain set of attributes, and the distance between words is calculated as the difference between the Cartesian product multiplied by the attributes of the word vectors. It is assumed that it is possible to select a pair of attributes of the same category for each pair of non-abstract words (e.g. the category could be *size*, and the distinctive attributed could be *big* and *small*). [Krebs and Paperno, 2016].

There is a certain amount of databases where words are associated with sets of different attributes. One of examples of such bases is a previously mentioned *BLESS* dataset which contains 200 pairs of words (for example, for the $[motorcycle, moped]$ word pair these are the two sets of attributes: $[large, small]$ and $[fast, slow]$) [Baroni and Lenci, 2011]. Another example is *Feature Norms Dataset* containing 24 963 pairs of words, for which a least one pair of distinctive features is selected (for example, for the pair $[airplane, helicopter]$ the *existence of wings* is selected) [Krebs and Paperno, 2016].

4.3.5. Semantic networks evaluation method uses manually constructed knowledge graphs (*semantic networks*) as a gold standard. In semantic networks, the words are organized in a graph in accordance with their semantic distinctive features based on judgments of teams of professional linguists. Semantic networks also feature a measure of similarity for word pairs based on the shortest path in a graph, so it could be compared with the similarity measure of the same pair calculated by word embeddings to evaluate its quality [Agirre et al., 2009].

The most well-known example of such semantic networks is the **WordNet**, a graph containing 155 287 words[5] organized in 117 659 synsets. [Hearst, 1998]. Another popular semantic network is **DBpedia**, a graph of concepts extracted from the Wikipedia and containing about 4.22 million words. Of course, there are even more different semantic networks, but they probably are less extensive then *WordNet* and *DBpedia*, so their representativeness in this task could be questioned.

4.4 Linguistic-driven methods

4.4.1. Phonosemantic analysis evaluation method is based on the idea that the form of a linguistic sign is not arbitrary, since it somehow correlates with its semantics. If that is true, it is possible to obtain certain data about semantics of a word through phonosemantic patterns of its phonemes or characters. In order to calculate phonosemantic difference between two words, one could measure use Levenshtein distance measure, and such metric could be used as a gold standard for evaluation [Gutiérrez et al., 2016]. Notably, this observation was confirmed not only for Latin alphabet, but also for Cyrillic [Kutuzov, 2017].

I am not aware of any open datasets designed specifically for evaluation on this task, and in each of the studies mentioned, the authors used their own data.

4.4.2. Bi-gram co-occurrence frequency evaluation method is based on the idea that the distance between the words vectors representing words of a phrase group (e.g. *noun + adjective*) should correlate with the frequency of this group in a corpus (bi-gram co-occurrence frequency). In other words, bi-gram co-occurrence frequency could be used as a gold standard [Kornai and Kracht, 2015].

I think that any representative corpus or dictionary of n-gram co-occurrence frequency dataset like the *Google 1T Frequency Dataset*[6] can be used for evaluation.

5 Future challenges

In this survey I attempted to systematize the existing attempts to explore and evaluate word embeddings. I highlighted existing problems in this field and described both mainstream and less well-known evaluation methods (16 intrinsic methods and 12 extrinsic methods, 28 evaluation methods overall). In conclusion I would like to briefly discuss the new problems that propose future challenges in the field of word embeddings evaluation.

The trends in the field of word embeddings are moving towards multi-language and multi-sense embeddings. Different approaches should be used for their evaluation, due to their different nature: in the case of multi-language

[5] On the moment of writing this work.

[6] https://books.google.com/ngrams

embeddings, there is one vector for translating a word to each of the supported languages, while in multi-sense embeddings, one word corresponds to multiple vectors, depending on the number of its senses. First attempts to investigate possible methods of evaluating such models are already made [Borbély et al., 2016,Reisinger and Mooney, 2010,Upadhyay et al., 2016], but I argue that mainstream approaches to evaluation like word similarity datasets would be even less applicable to such embeddings than to the "classic" mono-language and mono-sense embeddings since of strong differences in semantics of words in different languages.

There are also certain issues related to the nature of distributional word representations, that are not so implicitly related to the task of word embeddings evaluation. Among them are questions about distributional representations of compound linguistic units (phrases and sentences) [Lenci, 2017], interpretability of the vector components [Senel et al., 2017], connection with other models of formal and cognitive semantics [Lenci, 2008], etc.

Additionally, many studies dedicated to word embeddings evaluation miss one important factor which is a *bias of vector space* (I mean, bias in the terms of fairness) related to certain gender, racial or sexual orientation prejudices (for example, the similarity of gender-neutral words like *programmer* and the word *woman* should not be lower than the same similarity with the word *man*, but in some DSMs it is) [Bolukbasi et al., 2016,Garg et al., 2017]. Hence, if one considers that a "good" model should not be biased, then while evaluating the model she must take into account the robustness of the model to that bias. Due to this, the problems of evaluation and the problems of bias detection go hand in hand, and the complete solution of the first problem is impossible until the second one is solved.

Finally, I consider it an important problem that the success of resolving the task of word embeddings evaluation strongly depends on the existence of data; in other words, the task of evaluation is too supervised. Many researchers create a lot of materials and tools in English, and English embeddings can be evaluated very extensively, while for low-resource languages (like Urdu), even the simplest evaluation cannot be done. I believe it is important to make language-independent data for projecting models and data sets into a multi-language space in which the presence or absence of data for this language would not affect the ability to evaluate a distributional model.

Acknowledgements

I want to thank my colleagues, Andrey Kutuzov (University of Oslo) and Roman Suvorov (ISA RAS), for productive discussions on this paper, critique, suggestions and proofreading.

References

Abnar et al., 2017. Abnar, S., Ahmed, R., Mijnheer, M., and Zuidema, W. (2017). Experiential, distributional and dependency-based word embeddings have comple-