Russian Language Datasets in the Digital Humanities Domain and Their Evaluation with Word Embeddings

Gerhard Wohlgenannt*, Artemii Babushkin* and Denis Romashov* and Igor Ukrainets* and Anton Maskaykin* and Ilya Shutov*

*Faculty of Software Engineering and Computer Systems
ITMO University, St. Petersburg, Russia

Abstract. In this paper, we present Russian language datasets in the digital humanities domain for the evaluation of word embedding techniques or similar language modeling and feature learning algorithms. The datasets are split into two task types, word intrusion and word analogy, and contain 31362 task units in total. The characteristics of the tasks and datasets are that they build upon small, domain-specific corpora, and that the datasets contain a high number of named entities. The datasets were created manually for two fantasy novel book series ("A Song of Ice and Fire" and "Harry Potter"). We provide baseline evaluations with popular word embedding models trained on the book corpora for the given tasks, both for the Russian and English language versions of the datasets. Finally, we compare and analyze the results and discuss specifics of Russian language with regards to the problem setting.

Keywords: word embedding datasets, language model evaluation, Russian language, digital humanities, word analogy

1 Introduction

Distributional semantics base on the idea, that the meaning of a word can be estimated from its linguistic context [1]. Recently, with the work on word2vec [2], where prediction-based neural embedding models are trained on large corpora, word embedding models became very popular as input to solve many natural language processing (NLP) tasks. In word embedding models, terms are represented by low-dimensional, dense vectors of floating-point numbers. While distributional language models are well studied in the general domain when trained on large corpora, the situation is different regarding specialized domains, and term types such as *proper nouns*, which exhibit specific characteristics [3, 4]. Datasets in the digital humanities domain, which include some of these aspects, were presented by Wohlgenannt [5]. In this work, those datasets are translated to Russian language, and we provide baseline evaluations with popular word embedding models, and analyze differences between English and Russian experimental results. The manually created datasets contain *analogies* and *word* *intrusion* tasks for two popular fantasy novel book series: “A Song of Ice and Fire” (ASOIF, by GRR Martin) and “Harry Potter” (HP, by JK Rowling). The *analogy* task is a well-known method for the intrinsic evaluation of embedding models, the *word intruder* task is related to word similarity and used to solve the “odd one out” task [6].

The basic question is how well Russian language word embedding models are suited for solving such tasks, and what are the differences to English language datasets and corpora. More specifically, what is the performance on the two task types, which word embedding algorithms are more suitable for the tasks, and which factors are responsible for any differences between English and Russian language results?

In this work, we manually translated the datasets into Russian. In total, we present 8 datasets, for both book series, the two task types, and the distinction between unigrams and n-gram datasets. Word2vec [2] and FastText [7] with different settings were trained on the Russian (and English) book corpora, and then evaluated with the given datasets. The evaluation scores are sufficiently lower for Russian, but the application of lemmatization on the Russian corpora helps to partly close the gap. Other issues such as ambiguities in the translation of the datasets, and inconsistencies in the transliteration of English named entities are analyzed and discussed. As an example, the best accuracy scores for the ASOIF unigram dataset for Russian are 32.7% for the *analogy* task, and 73.3% for word intrusion, while for English the best results are 37.1%, and 86.5%, resp.

The main contributions include the eight Russian language datasets with 31362 task units in total, translated by two independent teams, baseline evaluations with various word2vec and FastText models, comparisons between English and Russian, and the analyses of the results, specifically with regards to corpus word frequency and typical issues in translation and transliteration.

The paper is structured as follows: After an overview of related work in Section 2, the two task types and the translated datasets are introduced in Section 3. Section 4 first elaborates the evaluation setup, for example the details of the corpora, and the model settings used in the evaluations. Subsequently, evaluation results, both aggregated and fire-grained, are presented. We discuss the findings in Section 5, and then provide conclusions in Section 6.

2 Related Work

Word embedding vectors are used in a many modern NLP applications to represent words, often by applying pre-trained models trained on large general-purpose text corpora. Ghannay et al. [8] compare the performance of model types such as word2vec CBOW and skip-gram [2], and GloVe [9]. For example, FastText [7] provides pre-trained models for many language for download. But there are also language-specific efforts, eg. RusVectōrēs [10] include a number of models trained on various Russian corpora. Workshops on Russian language semantic similarity [11] emphasized the importance of the research topic.

For the intrinsic evaluation of word embedding models, researchers often use existing word similarity datasets like WordSim-353 [12] or MEN [13], or analogy datasets like Google [2] or BATS [14].

In specialized domains, large text corpora for training are often not available. Sahlgren and Lenci [15] evaluate the impact of corpus size and term frequency on accuracy in word similarity tasks, and as expected, corpus size has a strong impact. The datasets that we translated contain a high percentage of named entities such as book characters and locations. Herbelot [3] discusses various aspects of instances (like named entities) versus kinds, such as the detection of instantiation relations in distributional models. Distributional models are shown to be better suited for categorizing than for distinguishing individuals and their properties [16]. This is also reflected by our results, esp. the analysis of task difficulty in word intrusion (see Section 4.2).

The work on using distributional methods in the digital humanities domain is limited. More efforts have been directed at dialog structure and social network extraction [17, 18] or character detection [19].

3 Tasks and Datasets

This section introduces the task types (analogy and word intrusion), discusses the dataset translation process, and briefly describes the word embedding algorithms used, as well as the basics of implementation.

3.1 Task Types

The datasets and evaluations focus on two task types: word intrusion and word analogy. Term analogy is a popular method for the evaluation of models of distributional semantics, applied for example in the original word2vec paper [2]. Word intrusion is a task similar to *word similarity*, which is a popular intrinsic evaluation method for word embedding models (see Section 2).

The word analogy task captures semantic or syntactic relations between words, a well-known example is "*man* is to *king*, like *woman* is to *queen*". The model is given the first three terms as input, and then has to come up with the solution (*queen*). Word embeddings models can, for example, apply simple linear vector arithmetic to solve the task, with $vector(man) - vector(woman) + vector(king)$. Then, the term closest (eg. measured by cosine similarity) to the resulting vector is the candidate term.

In the second task type, word intrusion, the goal is to find an intruding word in a list of words, which have a given characteristic. For example, find the intruder in: *Austria Spain Tokyo Russia*. In our task setup, the list always includes four terms, where one is the intruder to be detected.

3.2 Dataset Translation

The given datasets are based on extended versions of English language datasets presented in [5]. Those datasets were manually created inspired by categories and relations of online Wikis about "A Song of Ice and Fire" and "Harry Potter". The goal was to provide high quality datasets by filtering ambiguous and very-low frequency terms. The three dimensions (2 book series, 2 task types, unigram and n-gram datasets) led to eight published datasets.

In this work, the datasets were translated to Russian language. Two separate teams of native speakers translated the datasets to Russian. We found, that multiple book translations exist for the Harry Potter book series, and decided to work on two different book translations in this case. As many terms, esp. named entities like book characters and location names, have a slightly different translation or transliteration from the English original to Russian, we ended up with two independent datasets.

For ASOIF, both translation teams based their translations on the same Russian book version, and there where only slight differences – mostly regarding terms which are unigrams in English, but n-grams in Russian, and a few words which can have multiple translations into Russian.

3.3 Word Embedding Models

For the baseline evaluation of the datasets, we apply two popular word embedding models. Firstly, word2vec [2] uses a simple two-layer feed-forward network to create embeddings in an unsupervised way. The simple architecture facilitates training on large corpora. Basically, word similarity in vector space reflects similar contexts of words in the corpus. Depending on preprocessing, unigram or n-gram models can be trained. Word2vec includes two algorithms. CBOW predicts a given word from the window of surrounding words, while skip-gram (SG) predicts surrounding words from the current word. Secondly, FastText [7] is based on the skip-gram model, however, in contrast to word2vec, it makes use of sub-word information, and represents words as bag of character n-grams.

Hyperparameter tuning has a large impact on the performance of embedding models [15]. In the evaluations, we compare the results for different parameters settings for both datasets, details on those settings are found in Section 4.1.

3.4 Implementation

As mentioned, two teams worked independently on dataset translation and model training, which leads to the provision of two independent GitHub repositories[1][2]. The repositories can be used to reproduce the results, and to evaluate alternative methods – based on the book translation used. All library requirements, usage, and evaluation, and most importantly the datasets, are found in the repositories. For model creation and evaluation the popular Gensim library is used [6]. The implementation contains two main evaluation modules, one for the *analogies* task, and one for *word intrusion*. In the repository, word intrusion is coined *doesn't-match*, as this is the name of the respective Gensim function.

[1] `https://github.com/DenisRomashov/nlp2018_hp_asoif_rus`

[2] `https://github.com/ishutov/nlp2018_hp_asoif_rus`

Third parties can either reuse the provided evaluation scripts on a given embedding model, or use the datasets directly. The dataset format is the same as in word2vec [2] for *analogies*, and for the word intrusion task it is simple the understand, with the 4 words of the task unit, and the intruder marked.

4 Evaluation

4.1 Evaluation Setup

Book Corpora and Dataset Translation The models analyzed in the evaluations are trained on two popular fantasy novel corpora, “A Song of Ice and Fire” (ASOIF) by GRR Martin, and “Harry Potter” (HP) by JK Rowling.

From ASOIF, we took the first four books; the corpus size is 11.8 MB of plain text, with a total of 10.5M tokens (11.1M before preprocessing). The book series includes a large world with an immense number of characters, whereby about 30-40 main characters exists. Narration is mostly linear and the story is told in first person from the perspective of different main characters.

The HP book series consists of seven books, with a size of 10.7 MB and 9.2M tokens (9.8M before preprocessing). The books tell the story of young Harry Potter and his friends in a world full of magic. The complexity of the world, and the number of characters, is generally lower than in ASOIF.

The basics of dataset translations were already mentioned in Section 3.2. Two independent teams worked on the task. In case of ASOIF, both teams based their translation work and also model creation on the same Russian book corpus. For HP an original translation into Russian exists, which is still the most popular one. Later other translations emerged. In order to cover a wider range of corpora, and to investigate the differences between those translations, the teams used two different translations. The exact book versions are listed in the respective GitHub repositories[3][4].

Preprocessing In principle, we tried to keep corpus preprocessing to a minimum, and did only the following common steps: removal of punctuation symbols (except hyphens), removal of lines that contain no letters (page numbers, etc.), and sentence splitting. However, in comparison to the English original version of datasets and corpus, for Russian we found that term frequencies of the terms in the datasets were significantly lower. A substantial amount of dataset terms even fell below the `min_count` frequency used in model training and was thereby excluded from the models. This can be attributed to the rich morphology of Russian language, and other reasons elaborated in the discussion section (Section 5). For this reason, we created a second version of the corpora with lemmatization applied to all tokens of the book series[5]. In the evaluations, we present and compare results of both corpora versions, with and without lemmatization applied.

[3] `github.com/ishutov/nlp2018_hp_asoif_rus/blob/master/Results.md`

[4] `github.com/DenisRomashov/nlp2018_hp_asoif_rus/blob/master/RESULTS.md`

[5] Using this toolkit: `tech.yandex.ru/mystem`

Furthermore, for the creation of n-gram annotated corpora the *word2phrase* tool included in the word2vec toolkit [2] was utilized.

Models and Settings As mentioned, we train word2vec and FastText models on the book corpora – using the Gensim library. In the upcoming evaluations, we use the following algorithms and settings to train models:

w2v-default: This is a word2vec model trained with the Gensim default settings: 100-dim. vectors, word window size and minimum number of term occurrence are both set to 5, iter (number of epochs): 5, CBOW.

w2v-SG-hs : Defaults, except: 300dim. vectors, 15 iterations, number of negative samples: 0, with hierarchical softmax, with the skip gram method[6].

w2v-SG-hs-w12: like *w2v-SG-hs*, but with a word window of 12 words.

w2v-SG-ns-w12: like *w2v-SG-w12*, with negative sampling set to 15 instead of hierarchical softmax.

w2v-CBOW: like *w2v-SG-hs*, but with the CBOW method instead of skip-gram.

FastText-default: A FastText model trained with the Gensim default settings: those are basically the same default settings as for *w2v-default*, except for FastText-specific parameters.

FastText-SG-hs-w12: Defaults, except: 300dim. vectors, 15 iterations, a word-window of 12 words, number of negative samples: 0, with hierarchical softmax and the skip-gram method[7].

FastText-SG-ns-w12: like *ft-SG-hs-w12*, but with negative sampling (15 samples) instead of hierarchical softmax.

Table 1. Number of tasks and dataset sections (in parentheses) in the Russian language datasets – with the dimensions of task type, book corpus and unigram/n-gram

Book Corpus	Task Type	Unigram	N-Gram
HP	Analogies	4790 (17)	92 (7)
	Word Intrusion	8340 (19)	1920 (7)
ASOIF	Analogies	2848 (8)	192 (2)
	Word Intrusion	11180 (13)	2000 (7)

Datasets In total, we provide eight datasets. This number stems from three datasets dimensions: the task type (analogies and word intrusion, the two book series, and the distinction between unigram and n-gram datasets). Table 1 gives an overview of the number of tasks within the datasets, and also of the number of sections per task. Sections reflect a subtask with specific characteristics and

[6] size=300, -negative=0, sg=1, hs=1, iter=15

[7] size=300, -negative=0, sg=1, hs=1, iter=15, -window=12

difficulty, for example analogy relations between *husband* and *wife*, or between between a *creature* (individual) and its *species*. Typically, per section, the items on a given side of the relation are members of the same word or named entity category, therefore a distributional language model can be more deeply analyzed for its performance in those subtasks.

In contrast to popular word similarity datasets like WordSim-353 [12] or MEN [13], most of the dataset terms are named entities. Herbelot [3] studies some of the properties of named entities in distributional models. For example, in the unigram word intrusion dataset, only around 7% (ASOIF) and 17% (HP) of terms are *kinds*, the rest are named entities.

4.2 Evaluation Results

In this section, we present and analyze the evaluation results for the presented datasets using word embedding models. We start with an overview of the results of the *analogies* and *word intrusion* tasks, followed by more fine-grained results for the different subtasks of the *analogies* tasks. For the *word intrusion* task, we investigate evaluation results depending on task difficulty, and finally, a summary of results on n-gram datasets is presented. Further details on the results can be found on github[8][9].

Table 2. Overall *analogies* accuracy of the Russian unigram datasets for both book series. Results are given for models with and without lemmatization of the corpora. Values for English given in parenthesis.

Book series	ASOIF		HP	
Preprocessing	Minimal	Lemmatization	Minimal	Lemmatization
w2v-default	0.57 (8.15)	2.35 (-)	00.42 (6.88)	1.75 (-)
w2v-SG-hs	17.61 (28.44)	24.40 (-)	13.40 (25.11)	23.00 (-)
w2v-SG-hs-w12	**24.56** (37.11)	**32.66** (-)	**20.34** (30.00)	**28.95** (-)
w2v-SG-ns-w12	21.58 (29.32)	20.97 (-)	13.17 (20.84)	12.99 (-)
w2v-w12-CBOW	0.57 (2.67)	1.07 (-)	0.68 (7.22)	2.62 (-)
FastText-default	0.42 (1.33)	2.31 (-)	0.08 (0.87)	0.42 (-)
FastText-SG-hs-w12	11.04 (29.81)	21.58 (-)	8.57 (25.46)	19.30 (-)
FastText-SG-ns-w12	0.99 (14.64)	0.8 (-)	3.77 (14.23)	4.13 (-)

Table 2 provides an overview of results for the *analogies* task. It includes the results for the two book series ASOIF and HP. We distinguish two types of input corpora, namely with and without the application of lemmatization ("minimal" preprocessing vs. "lemmatization"). Embedding models were trained on the corpora with the settings described in Section 4.1. Furthermore, the evaluation scores for English language corpora and datasets are given for comparison.

[8] `github.com/ishutov/nlp2018_hp_asoif_rus/blob/master/Results.md`

[9] `github.com/DenisRomashov/nlp2018_hp_asoif_rus/blob/master/RESULTS.md`

The results in Table 2 indicate that models trained with the skip-gram algorithm clearly outperform CBOW for analogy relations. Another important fact is that for Russian language lemmatization of the corpus tokens before training has a strong and consistent positive impact on results. However, the numbers for Russian stay below the numbers for English. Both the performance impact of lemmatization, and the differences between English may partly be the result of differences in corpus term frequency. This intuition will be investigated and discussed in Section 5.

Table 3. Overall *word intrusion* accuracy (in percent) for the Russian unigram datasets for both book series. Results are given for models with and without lemmatization of the corpora. Values for English given in parenthesis.

Book series	ASOIF		HP	
Preprocessing	Minimal	Lemmatization	Minimal	Lemmatization
w2v-default	62.03 (86.53)	64.83 (-)	34.69 (64.83)	53.59 (-)
w2v-SG-hs	65.93 (77.9)	**73.30** (-)	55.99 (73.3)	60.44 (-)
w2v-SG-hs-w12	67.11 (74.86)	68.89 (-)	**61.09** (68.69)	59.87 (-)
w2v-SG-ns-w12	**68.09** (75.15)	67.1 (-)	58.43 (74.43)	57.01 (-)
w2v-w12-CBOW	57.35 (75.61)	61.28 (-)	42.39 (61.28)	48.73 (-)
FastText-default	61.59 (73.82)	56.56 (-)	41.92 (56.56)	46.50 (-)
FastText-SG-hs-w12	66.82 (75.99)	70.20 (-)	60.99 (70.2)	60.27 (-)
FastText-SG-ns-w12	67.81 (75.38)	68.41 (-)	59.13 (76.41)	**61.54** (-)
Stock Embeddings	27.3 (-)	- (-)	25.36 (-)	- (-)
Random Baseline	25.00	25.00	25.00	25.00

Table 3 gives the results for the *word intrusion* tasks. Again, we distinguish between preprocessing with and without lemmatization, and between the results of different models trained on the two book series. For the word intrusion task, the differences observed between skip-gram and CBOW, and regarding lemmatization, are smaller as compared to *analogies* results in Table 2; however, the tendencies still exist. For comparison, we also applied pretrained FastText models ("Stock Embeddings") trained on Wikipedia[10] to the task. As expected, those models perform very poorly, only slightly over the random baseline.

All datasets are split into various sections, which reflect specific relation types, for example *child-father* or *houses-and-their-seats*. Those relations have certain characteristics, such as involving person names, location entities, or other, which allow a fine-grained analysis and comparison of embedding models and their performance. Table 4 shows some selected sections from the ASOIF analogies dataset. The performance varies strongly over the different subtasks, but the data indicates, that models that do well in total, are also more suitable on the individual tasks.

[10] `https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md`

Table 4. ASOIF Analogies Russian dataset: Accuracy of different word embedding models on selected analogies task sections, and total accuracy. In parenthesis, values from the English language dataset are given for comparison.

Task Section	first-lastname	husband-wife	loc-type	houses-seats	Total
Number of tasks:	2368	30	168	30	2848
w2v-default	1.93 (8.78)	0.0 (6.67)	5.0 (4.76)	10.0 (20.0)	2.35 (8.15)
w2v-SG-hs	26.98 (32.01)	15.0 (10.0)	7.5 (11.9)	26.67 (40.0)	24.4 (28.44)
w2v-SG-hs-w12	**36.24** (42.36)	**20.0** (10.0)	6.25 (19.64)	**30.0** (33.33)	**32.66** (37.71)
w2v-SG-ns-w12	32.62 (40.62)	5.0 (6.67)	**12.5** (22.62)	26.67 (40.0)	29.62 (36.41)
w2v-w12-CBOW	0.69 (1.27)	5.0 (6.67)	2.5 (11.9)	6.67 (30.0)	1.07 (2.67)
FastText-default	2.33 (1.06)	5.0 (3.33)	0.0 (3.57)	3.33 (3.33)	2.31 (1.33)
FastText-SG-hs-w12	23.86 (34.04)	15.0 (6.67)	6.25 (13.69)	20.0 (46.67)	21.58 (29.81)
FastText-SG-ns-w12	27.33 (35.09)	15.0 (3.33)	5.0 (11.9)	13.33 (26.67)	24.4 (30.44)

Table 5. Accuracy results with regards to task difficulty – Russian ASOIF word intrusion dataset (unigrams), trained on a lemmatized corpus.

Task Difficulty	1 (hard)	2 (med-hard)	3 (medium)	4 (easy)	AVG
Number of tasks:	2795	2795	2795	2795	11180
w2v-default	61.82	**72.9**	70.16	91.74	**74.17**
w2v-SG-hs	46.27	67.72	73.38	85.33	68.18
w2v-SG-hs-w12	39.18	67.73	**75.67**	85.83	67.1
w2v-SG-ns-w12	57.53	70.98	74.35	90.84	73.43
FastText-default	**61.86**	65.62	66.26	**96.85**	72.65
FastText-SG-hs-w12	46.76	67.59	73.38	87.48	68.8
FastText-SG-ns-w12	39.28	66.87	73.56	86.37	66.52

The word intrusion datasets were created with the idea of four task difficulty levels. The *hard* level includes near misses; on the *medium-hard* level, the outlier still has some semantic relation to the target terms, and is of the same word (or NE) category. On the *medium* level outliers are of the same word category, but have little semantic relatedness to the target term. And finally, in the *easy* category, the terms have no specific relation to the target terms. As an example, if the target terms are `Karstark Greyjoy Lannister`, ie. names of *houses*, then a *hard* intruder might be `Theon`, who is a person from one of the houses. `Bronn` will be a *med-hard* intruder, also a person, not from those houses. `Winterfell` (a location name) will be in the *medium* category, and `raven` in the easy one.

Very interestingly, models using the CBOW algorithm (such as *w2v-default* and *FastText-default*) provide very good results on the hardest task category with over 60% of correct intruders selected. On the other hand, SG-based models only show 39%-46%. For the easiest category, *FastText-default* excels with almost

97% accuracy. In general word embeddings, esp. when trained on small datasets and using cosine similarity for the task, struggle to single out terms in the *hard* category by a specific (minor) characteristic of the target terms. This will be further discussed in section 5.

N-Gram Results As mentioned, in addition to the four unigram dataset, we created complementary n-gram datasets for the two book series and the two task types. In correspondence with the n-gram detection method used [2], in the datasets n-grams are words connected by the underscore symbol. Many of the terms are person or location names such as *Forbidden_ Forest* or *Maester_ Aemon*. For reasons of brevity, we will not include result tables here (see GitHub for details). The general tendency is that n-gram results are below unigram results in the *analogies* task, for word intrusion results are comparable with around 70% accuracy (depending on model settings). In comparison with word2vec, FastText-based models perform better on n-gram than unigram tasks, this can be explained by the capability of FastText to leverage subword-information within n-grams.

5 Discussion

In general, there is quite a big difference in performance between Russian and English datasets and models, when the same (minimal) preprocessing is being applied to the corpora. For example, in Table 2 the best ASOIF performance for Russian (with minimal preprocessing) is 24.56%, but 37.11% for English, and for HP the values are 20.34% for Russian, and 30.00% for English. In the case of *word intrusion*, the same pattern repeats: 68.09% for Russian ASOIF, and 86.53% for English, and finally, 61.09% for the Russian HP dataset versus 74.43% (English).

Our first intuition was, that the rich morphology of the Russian language, where also proper nouns have grammatical inflections by case, might reduce the frequency of dataset words in the corpora. Sahlgren and Lenci [15] show the impact of term frequency on task accuracy in the general domain. Subsequent analysis shows, that eg. for the HP dataset, the average term frequency of dataset terms is 410 for the English terms and English book corpus, while for the Russian it is only 249. We then decided to apply lemmatization to the Russian corpora, which helped to raise Russian average term frequency to 397. Also the evaluation results improved overall, as seen in the tables in Section 4.2. However, despite the positive effects, lemmatization also introduces a source of errors. For some dataset terms, the frequency even becomes lower; after lemmatization, the number of Russian dataset terms that are below the *min_ count* threshold to be introduced into the word embedding models rises. An example of such problem cases is the word “Fluffy” from HP, which was translated as “Пушок”. But then the lemmatizer wrongly changed it to “пушка” (a gun), so that “Пушок” disappeared from the trained models.

A number of other difficulties and reasons for the lower performance on the Russian datasets emerged: a) When comparing the output of the two translation teams, we observed words that have many meaningful translations to Russian, thereby lowering term frequency. For example, “intelligence” can be translated as “ум”, “интеллект”, “осознание”, “остроумие” and so on. b) The transliteration of English words into Russian is not always clear, and we have found in the analysis that even within the same book corpora (translations) it is not always consistent, even more so between different translators. For example, there is no [æ] phonetic sound in Russian, so it can be transliterated to multiple letters: *a*, *e*, *э*. c) In Russian, the *ё* letter is often replaced with *e*. If this happens inconsistently, it impacts term frequencies.

Another interesting aspect is the performance of the models on various difficulty levels in the word intrusion tasks. If difficulty is low, then common word embedding models already work very well, in our experiments with an accuracy up to 97%. However, in the *hard* category terms are very similar in their overall semantics and context, but the target terms possess one characteristic that the intruder lacks. Cosine similarity just looks at overall vicinity in vector space, with a success rate of ca. 40–60%. There has been some work on named entities and the distinction between *individuals* (and their properties) and *kinds* within distributional models [3, 16, 4], but there is still much room on how to tackle such issues in a general way.

6 Conclusions

In this work, we present Russian language datasets in the digital humanities domain for the evaluation of distributional semantics models. The datasets cover two basic task types, *analogy* relations and *word intrusion* for two well-known fantasy novel book series. The provided baseline evaluations with word2vec and FastText models show that models for the Russian versions of the corpora and datasets offer lower accuracy than for the English originals. The contributions of the work include: a) the translation to Russian and provision (on GitHub) of eight datasets in the digital humanities domain, b) providing baseline evaluations and comparisons for various settings of popular word embedding models, c) studying the effects of preprocessing (esp. lemmatization) on performance, d) analyzing the reasons for differences between Russian and English language evaluations, most notably term frequency and issues arising from translation.

7 Acknowledgments

This work was supported by the Government of the Russian Federation (Grant 074-U01) through the ITMO Fellowship and Professorship Program.

References

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

2. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
3. Herbelot, A.: Mr darcy and mr toad, gentlemen: distributional names and their kinds. In: Proc. 11th Int. Conf. on Computational Semantics. (2015) 151–161
4. Boleda, G., Padó, S., Gupta, A.: Instances and concepts in distributional space. In: EACL 2017, Valencia, Spain, April 3-7, 2017, Vol. 2: Short Papers. (2017) 79–85
5. Wohlgenannt, G., Chernyak, E., Ilvovsky, D., Barinova, A., Mouromtsev, D.: Relation extraction datasets in the digital humanities domain and their evaluation with word embeddings. In: CICLING 2018. Volume upcoming of Lecture Notes in Computer Science., Hanoi, Vietnam, Springer (March 2018) upcoming
6. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, ELRA (May 2010) 45–50
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
8. Ghannay, S., Favre, B., Estève, Y., Camelin, N.: Word embedding evaluation and combination. In et al., N.C., ed.: Proc. of the LREC 2016, Paris, France, ELRA (may 2016)
9. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). (2014) 1532–1543
10. Kutuzov, A., Kuzmenko, E. In: WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. Springer International Publishing, Cham (2017) 155–161
11. Panchenko, A., Loukachevitch, N.V., Ustalov, D., Paperno, D., Meyer, C., Konstantinova, N.: Russe: The first workshop on russian semantic similarity. CoRR **abs/1803.05820** (2015)
12. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: The concept revisited. In: Proc. of the 10th int. conf. of the World Wide Web, ACM (2001) 406–414
13. Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. J. Artif. Int. Res. **49**(1) (January 2014) 1–47
14. Gladkova, A., Drozd, A., Matsuoka, S.: Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In: SRW@HLT-NAACL, ACL (2016) 8–15
15. Sahlgren, M., Lenci, A.: The effects of data size and frequency range on distributional semantic models. In: EMNLP, ACL (2016) 975–980
16. Boleda, G., Padó, S., Pham, N.T., Baroni, M.: Living a discrete life in a continuous world: Reference in cross-modal entity tracking. In: IWCS 2017 – 12th International Conference on Computational Semantics — Short papers. (2017)
17. Elson, D.K., Dames, N., McKeown, K.R.: Extracting social networks from literary fiction. In: Proc. 48th Annual Meeting of the ACL. ACL '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 138–147
18. Jayannavar, P., Agarwal, A., Ju, M., Rambow, O.: Validating literary theories using automatic social network extraction. In: CLfL@ NAACL-HLT. (2015) 32–41
19. Vala, H., Jurgens, D., Piper, A., Ruths, D.: Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In Màrquez, L.e.a., ed.: EMNLP, ACL (2015) 769–774