Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4227–4234
Marseille, 20-25 June 2022

IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set

Steinunn Rut Friðriksdóttir, Hjalti Daníelsson,
Steinþór Steingrímsson, Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavik, Iceland
srf2@hi.is, hjalti.danielsson@arnastofnun.is,
steinthor.steingrimsson@arnastofnun.is, einar.freyr.sigurdsson@arnastofnun.is

Abstract

Word embedding models have become commonplace in a wide range of NLP applications. In order to train and use the best possible models, accurate evaluation is needed. For extrinsic evaluation of word embedding models, analogy evaluation sets have been shown to be a good quality estimator. We introduce an Icelandic adaptation of a large analogy dataset, BATS, evaluate it on three different word embedding models and show that our evaluation set is apt at measuring the capabilities of such models.

Keywords: word embeddings, Icelandic, BATS, analogy test set

1. Introduction

Over the past decade, there has been significant development of vector space semantics in computational linguistics. Mikolov et al. (2013b) demonstrated how the vector offset method can be used to capture syntactic and semantic regularities using word embeddings. However, in order to show whether a language model can successfully perform analogical reasoning, it is necessary to showcase which types of relations it can handle. Gladkova et al. (2016) presented a new test set that measures language models' abilities to recognize these linguistic regularities in a balanced and comprehensive way. The Bigger Analogy Test Set (BATS) consists of 98,000 proportional analogies, that is to say, questions in the form of a:b::c:d. The questions are balanced across 4 types of relations in 40 subcategories: inflectional and derivational morphology and lexicographic and encyclopedic semantics. Each subcategory includes 50 word pairs representing a specific type of linguistic relation, e.g. synonymy, plurality, suffixation etc.

There are three aspects that separate BATS from previous popular analogy test sets such as the *Google analogy test set* (Mikolov et al., 2013a), often considered as a benchmark for word embeddings. Firstly, unlike the Google analogy test set which contains 9 morphological categories (most of which are inflectional) and 5 semantic categories (over half of which are of the category country::capital), BATS aims to be well balanced in the types of linguistic relations it contains. The categories of inflectional and derivational morphology, and lexicographic and encyclopedic semantics, are all equally represented, each containing 10 subcategories with 50 word pairs each. Secondly, the morphological categories have been sampled to reduce homonymy. This means that words that can be of more than one grammatical category, depending on context (e.g. *walk* which can either be a noun or a verb), are avoided. There is, however, some ambiguity in the semantic categories as they tend to be smaller and their word candidates often have multiple functions. In order to properly avoid homonymy in these categories, it would be necessary to include primarily infrequent words which is not practical for testing purposes. Thirdly, BATS offers multiple correct answers to the semantic analogy questions where applicable. An obvious example of this would be that each hypernym has multiple hyponyms and vice versa.

In recent years, there has been significant development in Icelandic language technology, particularly concerning data collection and building the foundation for future development. The Icelandic language technology project plan for 2018–2022 (Nikulásdóttir et al., 2017) identifies five core projects, some of which involve word embeddings in one way or another. As previously stated, the vector offset method works surprisingly well to mirror linguistic relations and one way to measure the embeddings' performance is to have them perform analogical reasoning. No analogy test set has been made previously for Icelandic word embeddings. We thus present IceBATS (Friðriksdóttir et al., 2021), a new Icelandic analogy test set which is comparable to BATS in every major way. In Sections 2.2–2.6, we discuss the minimal yet necessary changes we have made compared to the English dataset. In Section 3 we describe how we trained different word embeddings and tested them on IceBATS.

2. Data and Candidate Selection

For all intents and purposes, the categories of IceBATS are the same as in the original set. Like BATS, IceBATS is divided into inflectional and derivational morphology, and encyclopedic and lexicographic semantics, each containing 10 subcategories of 50 word pairs. This yields 98,000 unique analogy questions.

2.1. The Corpus

In order to gain an understanding of word frequency in Icelandic (and to train our models, discussed in Section 3), we look to the Icelandic Gigaword Corpus, hereafter referred to as the IGC (Steingrímsson et al., 2018). It is by far the largest text corpus available in Icelandic and has been in constant expansion since its original publication in 2018. The 20.05 version (Steingrímsson and Barkarson, 2020) contains approximately 1.5 billion running words of text, each tagged with morphosyntactic information and accompanied by its lemma. The lemmatization and tagging are automated and not manually corrected. The corpus contains various types of text, including official texts (such as parliamentary speeches), news articles and various texts from the text collection of the Árni Magnússon Institute for Icelandic Studies. It is therefore a valuable source of information on Icelandic language and can be used to estimate overall word frequencies. We used IGC as a reference point for our decision-making when compiling IceBATS.

2.2. Frequency Thresholds and Proportions

It is important that the majority of words used in IceBATS are neither of very low nor very high frequency. Words that are very uncommon are unlikely to appear in the corpus at all which can potentially create a false negative result where the test is simply too hard for a model to do well on. On the other hand, testing only very common words can create false confidence in the model's performance. IceBATS has 4,135 unique words. 624 words (15.1%) appear 0–100 times in our corpus, 629 words (15.2%) appear 101–500 times, 404 words (9.7%) appear 501–1,000 times, 1,457 words (35.2%) appear 1,001–10,000 times and 1,021 words (24.7%) appear over 10,000 times in the corpus. We strive to keep the majority of our word candidates in the middle section. As IGC is in constant expansion, these numbers might of course change somewhat in future versions or be less applicable to models trained on other corpora. However, we still consider it reflective of actual use of the Icelandic language and should therefore serve its purpose as a reference point.

2.3. Inflectional Morphology

Icelandic is a morphologically rich language which requires a slightly different approach than English when it comes to the category of inflectional morphology. Just like the original BATS, IceBATS considered three main parts of speech: **nouns**, **verbs** and **adjectives**. However, while the original dataset heavily favors verbs (6 out of 10 subcategories), we decided to preserve balance while slightly favoring nouns as they are by far the most frequent word class out of the three, a claim backed up by counting the POS tags of our corpus of choice where nouns are approximately 78% of all words. Additionally, we need to consider defining characteristics of Icelandic nouns: the four cases, three genders and the suffixed article. For the predictability of the declension of nouns in Icelandic, the nominative singular, genitive singular and the nominative plural are generally the most important forms. We therefore put the strongest emphasis on these as seen in the subcategories singular::plural nominative and nominative::genitive singular. The other two subcategories are indefinite form::definite article singular and indefinite form::definite article plural. Our research of the corpus indicates that the three genders are spread relatively evenly, the feminine being the most common and the neuter the least common by a small margin. We divide all subcategories into the three genders, usually (but not always) slightly favoring the feminine. Approximately 75% of the nouns in our corpus have a strong declension so we favor them as well in our selection.

The original BATS has two subcategories containing adjectives where the positive degree is compared to the comparative and the superlative, respectively. As Icelandic adjectives get a gender value for declension from the noun they modify and due to the fact that the superlative is the least common degree of adjectives in our corpus, we decided to include only the positive::comparative subcategory but for all three genders separately. Approximately 61% of the adjectives in our corpus have an indefinite ("strong") declension and 7% are indeclinable (making them unsuitable for IceBATS). We keep the division in our subcategories similar, with 70–80% of the selected adjectives having strong declension and 20–30% having weak declension.

As for verbs, IceBATS follows the precedence set by the original BATS and compares the principal parts. There are three different types of verbs in IceBATS with respect to inflection: the ones with weak inflection, the ones with strong inflection and preterite-present verbs. These do not all have exactly the same principal parts: weakly inflected verbs have three whereas the other two types of verb have four. We thus compare only the principal parts that apply to all different types: the infinitive, the first person singular indicative past tense, and the past participle (making up the following categories: infinitive::indicative singular; infinitive::past participle; past participle::indicative singular). According to our research, approximately 45% of the verbs in our corpus have a strong inflection, 54% have a weak inflection and the remainder consists of preterite-present verbs. We strive to keep similar proportions where 4 out of 50 word pairs consist of preterite-present verbs in all three subcategories but the proportions of strong and weak declension vary slightly.

An example of each subcategory of inflectional morphology is shown in Table 1.

2.4. Derivational Morphology

Naturally, some of the largest variation between the original BATS and IceBATS is found in the deriva-

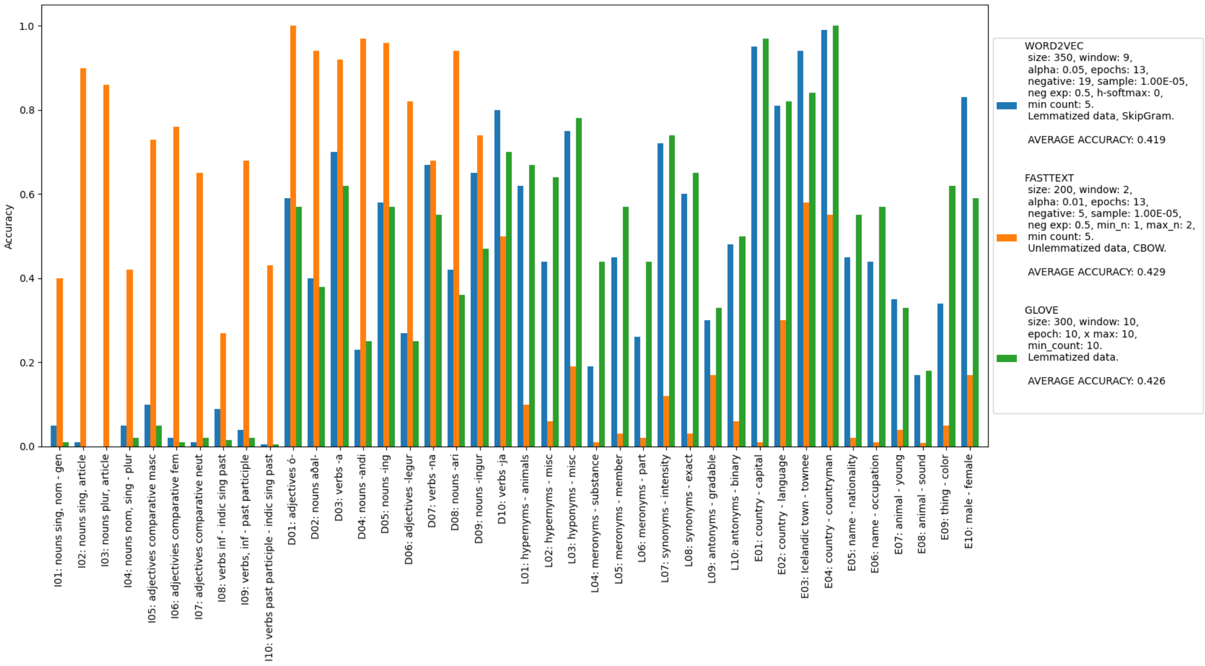

Figure 1: IceBATS performance of three different word embedding models, trained with various hyperparameter settings.

I01	**nouns singular, nominative – genitive** veður – veðurs vegur – vegar
I02	**nouns singular, article** krafa – krafan veður – veðrið
I03	**nouns plural, article** kröfur – kröfurnar embætti – embættin
I04	**nouns nominative, singular – plural** vegur – vegir félag – félög
I05	**adjectives comparative masculine** brúnn – brúnni veikur – veikari
I06	**adjectives comparative feminine** merkileg – merkilegri grunn – grynnri
I07	**adjectives comparative neuter** ljóst – ljósara gott – betra
I08	**verbs, infinitive – indicative sing. past** fara – fór sofa – svaf
I09	**verbs, infinitive – past participle** syngja – sungið fara – farið
I10	**verbs, past participle – ind. sing. past** gert – gerði sagt – sagði

Table 1: Inflectional morphology: subcategories and examples.

tional morphology top category. The original BATS divides the subcategories into word pairs where the stem does not change and word pairs where the stem changes. In essence, we preserve this division but morphology is rarely as clean cut in Icelandic as it is in English. Three of our subcategories contain word pairs where the only change to the word is the added affix. These are: adjectives with the prefix *ó-* 'un' (*skemmtilegur* 'fun' – *óskemmtilegur* 'not fun', *ó-* reverses or negates the meaning); nouns with the prefix *aðal-* (*leikkona* 'actress' – *aðalleikona* 'main actress', *aðal-* implies primary importance); and verbs with the suffix *-a* (*hopp* 'jump (noun)' – *hoppa* 'jump (verb)'). Three of our subcategories make use of a suffix that is added to the stem. These are: nouns with the suffix *-andi* (*eiga* 'own' – *eigandi* 'owner', *-andi* is a nominalizer and creates an agent nominal); nouns with the suffix *-ing* (*dreifa* 'scatter' – *dreifing* 'dispersion', *-ing* is a nominalizer); and adjectives with the suffix *-leg(ur)* (*nauðsyn* 'necessity' – *nauðsynlegur* 'necessary', *-leg(ur)* is an adjectivizer). The remaining subcategories have stem changes (usually a sound shift) in at least some of their word pairs: verbs with the suffix *-na* (*blautur* 'wet' – *blotna* 'get wet'); nouns with the agent nominal suffix *-ari* (*dæma* 'judge' (verb) – *dómari* 'judge (noun)', *-ari* is an agent nominalizer); nouns with the agent nominal suffix *-ing(ur)* (*andstaða* 'opposition' – *andstæðingur* 'opponent'); and verbs with the suffix *-ja* (*glaður* 'glad' – *gleðja* 'please').

An example of each subcategory is shown in Table 2.

2.5. Encyclopedic Semantics

In Gladkova et al. (2016), the original BATS team explains that the encyclopedic semantics category is based on Wikipedia word lists and other internet resources along with the color dataset (Bruni et al., 2012) and the Google analogy test set. IceBATS fol-

D01	**adjectives ó-** breyttur – óbreyttur ábyrgur – óábyrgur
D02	**nouns aðal-** atriði – aðalatriði áhersla – aðaláhersla
D03	**verbs -a** mynd – mynda hopp – hoppa
D04	**nouns -andi** eiga – eigandi stjórna – stjórnandi
D05	**nouns -ing** dreifa – dreifing alhæfa – alhæfing
D06	**adjectives -legur** nauðsyn – nauðsynlegur alvara – alvarlegur
D07	**verbs -na** blautur – blotna harður – harðna
D08	**nouns -ari** dæma – dómari kenna – kennari
D09	**nouns -ingur** vík – víkingur andstaða – andstæðingur
D10	**verbs -ja** sagður – segja glaður – gleðja

Table 2: Derivational morphology: subcategories and examples.

lows their example almost exactly, with the exclusion of the subcategory animals::shelter (e.g. *fox – den*) which has been replaced with the subcategory country::countryman (e.g. *Columbia – Columbian*). This is mostly due to the fact that animal shelters all tend to be named very similarly in Icelandic and/or are of very low frequency in the corpus. Additionally, the original subcategory of UK city::county (e.g. *York – Yorkshire*) has been replaced by one of Icelandic towns and the names of their inhabitants (e.g. *Reykjavík – Reykvíkingur*). Both the original and IceBATS are therefore divided into geography-related subcategories (country::countryman, Icelandic town::townee, country::capital, country::language), people-related subcategories (name::nationality, name::occupation), animal-related subcategories (animal::sound, animal::young) and two miscellaneous subcategories (thing::color and male::female). This means that IceBATS is slightly more geography-heavy than the original BATS. This does not seem to affect our results significantly.

An example of each subcategory of encyclopedic semantics is shown in Table 3.

E01	**country – capital** Ísland – Reykjavík Danmörk – Kaupmannahöfn
E02	**country – language** Danmörk – danska Frakkland – franska
E03	**Icelandic town – townee** Ísafjörður – Ísfirðingur Borgarnes – Borgnesingur
E04	**country – countryman** Bandaríkin – Bandaríkjamaður England – Englendingur
E05	**name – nationality** Aristóteles – grískur/forngrískur Lenín – rússneskur/sovéskur
E06	**name – occupation** Björk – söngkona/tónlistarkona/ tónlistarmaður/popptónlistarmaður Aristóteles – heimspekingur/kennari/ vísindamaður
E07	**animal – young** kýr – kálfur/kvíga/kvígukálfur/ungkýr/ ungneyti/ungnaut kind – lamb/lambhrútur/gimbur/gemlingur
E08	**animal – sound** kýr – baula hundur – gelta/urra/ýlfra/væla/ spangóla/gjamma
E09	**thing – color** gras – grænn/grasgrænn/gulur/ brúnn/gulgrænn banani – gulur/brúnn/grænn
E10	**male – female** karl – kona bróðir – systir

Table 3: Encyclopedic semantics: subcategories and examples.

2.6. Lexicographic Semantics

The lexicographic semantics category of the original BATS is based on SemEval2012-Task 2 (Jurgens et al., 2012) as well as BLESS (Baroni and Lenci, 2011) and EVALution (Santus et al., 2015). IceBATS includes exactly the same subcategories as the original BATS here and for the most part, the word pairs are translations of the original. This however varies a bit depending on the frequency threshold discussed in Section 2.2. We explore five different types of linguistic relations. Three subcategories deal with meronyms: part (e.g. *engine – car*), substance (e.g. *water – sea*) and member (e.g. *player – team*). Two subcategories are made of antonyms: binary (e.g. *black – white*) and gradable (e.g. *big – small/tiny/petite*), and of synonyms: binary (e.g. *sofa – couch*) and intensity-related (e.g. *cry – whine/weep/scream*). Two categories deal with hypernyms: miscellaneous (e.g. *plum – fruit*, *shirt – clothes*) and animals-related (e.g. *cat – feline*). The last category covers miscellaneous hyponyms (e.g. *bag – pouch*, *color – white*).

An example of each subcategory is shown in Table 4.

3. Testing Word Embeddings

In order to evaluate our test set, we trained various word embedding models on IGC and tried to optimize their performance on IceBATS using different hyperparameters, see Figure 1. Just like the original BATS team, we evaluate the models using the vector offset method (Mikolov et al., 2013a). The method computes

L01	**hypernyms – animals** kýr – húsdýr/spendýr/jórturdýr/seildýr/klaufdýr/slíðurhyrningur humar – krabbadýr/sjávardýr/sjávarfang/botndýr/fiskur/liðdýr/stórkrabbi
L02	**hypernyms – miscellaneous** tölva – tæki/tækni/rafmagnstæki/miðill kjóll – fatnaður/föt/klæði/klæðnaður/fatakostur
L03	**hyponyms – miscellaneous** poki – bakpoki/innkaupapoki/plastpoki/maíspoki/pappírspoki/burðarpoki/bréfpoki/flöskupoki/flögupoki/snakkpoki/nammipoki bók – ævintýri/vísindaskáldskapur/ævisaga/spennusaga/fræðibók/krimmi/glæpasaga/ástarsaga/riddarasaga/Íslendingasaga/fornsaga/drama
L04	**meronyms – substance** skegg – hár flaska – gler/plast
L05	**meronyms – member** hreindýr – hjörð/hópur lag – plata/diskur/tónverk/vínilplata/geisladiskur
L06	**meronyms – part** svefnherbergi – íbúð/hús/bygging/hótel/heimili fingur – hönd
L07	**synonyms – intensity** hræddur – óttasleginn/smeykur/uggandi/skelfdur/felmtraður/skelkaður/dauðhræddur/lafhræddur/skíthræddur reiður – byrstur/gramur/heiftugur/illur/bálreiður/öskuvondur/ofsareiður
L08	**synonyms – exact** tungl – máni vondur – illur/slæmur
L09	**antonyms – gradable** ódýr – dýr/verðmætur/ómetanlegur bjartur – dökkur/dimmur/daufur/litlaus
L10	**antonyms – binary** hvítur – svartur uppi – niðri

Table 4: Lexicographic semantices: subcategories and examples.

cosine similarity between a simple mean of the projection weight vectors and all other keys in the model. The idea is that linguistic relations are reflected in the distance between vectors so that similar relations are represented by similar vector distances between their elements. If vector A shares the same type of relation to vector B as vector C does to vector D, the distance between A and B should be nearly the same as the distance between C and D. Our code is heavily influenced by Gensim (Řehůřek and Sojka, 2010) and uses some of their code unchanged. The biggest difference, as inspired by the original BATS team, is that we allow for multiple correct answers where applicable in our analogy questions. The model is thus not faulted for a wrong guess unless none of the available answers is predicted as the correct one. Additionally, we have made a significant modification that deviates from the original BATS team in that we do not exclude the words A, B and C from the vocabulary before predicting the answer D. We note that in many cases the relations between those words, rather than simply being one-to-one, may form a hierarchy of one-to-many. This can in turn lead to a natural and perfectly acceptable occurrence where A and C have a sufficiently similar type of relationship with B so that B and D might be considered one and the same, such as in the question: "Britain is to English as the USA is to what?", or where A and C form such a natural subset of B that D can naturally be an equivalent set, such as in: "Britain is to Europe as Iceland is to what?" This change makes a minimal improvement to the results, raising the semantic categories by less than 1%.

For our research, we trained three types of models, word2vec (Mikolov et al., 2013a), FastText (Bojanowski et al., 2017) and GloVe (Pennington et al., 2014). The former two were also trained using both Skipgram (SG) and continuous bag of words (CBOW). Our average scores are almost twice as high as the ones from the original BATS team, with average accuracy between 0.419 and 0.429, depending on our models, against 0.221 to 0.285 in the models the original BATS team evaluated. A significant difference between our research and the original is that we tested our models on both lemmatized and nonlemmatized data. The models trained on lemmatized data test consistently higher on the semantic categories whereas it is necessary to use the nonlemmatized word forms to achieve good results for the morphological categories. We also find that CBOW models work better than Skipgram when training word2vec and fastText with the aim of getting a high score on the morphological categories, and that it is not necessary to have as high dimensional models when training for that purpose. For our data the accuracy gain is negligent when we pass 200 dimensions, as compared to 350 dimensions for the inflectional categories. When training models for semantic use, Skip-

Models and hyperparameters						
	Word2Vec		**FastText**		**GloVe**	
hyperparam.	**Infl. score**	**Encycl. score**	**Infl. score**	**Encycl. score**	**Infl. score**	**Encycl. score**
Architecture	CBOW	SkipGram	CBOW	SkipGram		
Epoch	13	20	13	45	20	20
Lemmatized	No	Yes	No	Yes	No	Yes
Dimensions	200	350	200	350	350	350
Window Size	1	8	2	5	10	10
Alpha	0.02	0.05	0.01	0.025		
Neg. sampling	13	19	5	19		
Sample	0.0001	0.00001	0.00001	0.00001		
Neg. exp.	0.2	0.5	0.5	0.5		
Min. word count	5	5	5	5	10	10
X max					40	40
Score	0.76	0.84	0.64	0.73	0.72	0.73

Table 5: Hyperparameters for our highest scoring models in two types of categories, inflectional morphology and encyclopedic semantics. All models were trained on IGC 20.05, a corpus of approximately 1.5 billion words. Further tuning may yield even better results.

gram also gives more accurate results. After training for more than 20 epochs, the accuracy gain with training further is very small, and for the morphological categories, training a CBOW model with 200 dimensions for more than 13 epochs is detrimental for our results. Lowering the window size will improve results on the morphological categories when using word2vec and fastText but keeping it relatively high will improve results when using GloVe. Likewise, keeping the window size at the high end will produce better results on the semantic categories. Lowering the alpha parameter, the initial learning rate in the model training, results in improvement for the morphological categories, but keeping it at close to 0.05 will be better for the semantic ones. In short, it is apparent that no one model with one set of parameters will be best suited for all four categories but it is possible to optimize a morphologically smart model on the one hand and a semantically smart model on the other. Our highest scoring models in the inflectional and encyclopedic categories are described in Table 5.

Interestingly, our results for derivational morphology are much higher than those of the original BATS team and that applies for both lemmatized and unlemmatized data. As expected, the subcategories containing sound shift in the word stem score slightly lower than the others. Despite this, all of our derivational subcategories score higher than their English equivalents. The inflectional morphology category shows similar but slightly lower results when compared to the English equivalent which is not surprising considering the complexity of Icelandic inflections. Our scores for the semantic categories are overall higher than those of the original BATS team when the models are trained on lemmatized data, indicating that lemmatization plays an important role in filtering out noise from the results.

In their paper, the original BATS team talks about the importance of word frequency as the categories containing more lower-frequency word pairs scored lower than those with higher-frequency word pairs. As explained in Section 2.2, we made sure to balance out the frequency of our candidates and this could in part explain why our results are higher. Additionally, we only cut off words that appear less than 5 times in the corpus but the original BATS team cut off words appearing less than 100 times, resulting in a larger vocabulary in our case.

4. Related Work

The quality of word embeddings is typically assessed with analogy tests. The developers of word2vec introduced the Google analogy test set, but other tests such as SimLex-999 (Hill et al., 2015) and WordSim-353 (Finkelstein et al., 2002) have also been popular. While these test sets have all been published in English, for some other languages few or none exist. Leviant and Reichart (2015) created multilingual versions of SimLex-999 and WordSim-353, with German, Italian and Russian as well as English. BATS, the analogy set we adapt for Icelandic, has previously been adapted to Japanese (Karpinska et al., 2018). A Portuguese analogy test set created from scratch, TALES (Oliveira et al., 2020), bears some resemblance to BATS in the way that it covers different types of lexical-semantic relations and has the same number of entries, 50.

5. Availability and Licensing

IceBATS (Friðriksdóttir et al., 2021) has been made available at the Icelandic CLARIN repository under a CC BY 4.0 license[1]. Word embeddings trained and evaluated on IceBATS as described in this paper are also available with accompanying metadata,

[1] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/120

word2vec models (Friðriksdóttir et al., 2022c), GloVe models (Friðriksdóttir et al., 2022b) and FastText models (Friðriksdóttir et al., 2022a). Code used for training and evaluation is on GitHub [2], and all information on the data at a website dedicated to word embeddings and evaluation datasets for Icelandic[3].

6. Conclusion

Language modelling is fundamental to natural language processing and word embeddings have become ever more important. In order to properly utilise them to their fullest potential, there has to be a comprehensive way to evaluate their performance. In this paper, we have discussed our adaptation of BATS and our accompanying research on various word embeddings' performance. No such test has previously been available for Icelandic. IceBATS offers an extensive overview of various linguistic relations and evaluates how well they are captured by the vector offset method. We have trained three different types of word embeddings and tested them accordingly. Our results show substantial improvement when compared to their English equivalent.

7. Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019–2023. The programme, which is managed and coordinated by Almannarómur[4], is funded by the Icelandic Ministry of Education, Science and Culture.

8. Bibliographical References

Baroni, M. and Lenci, A. (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20:116–131.

Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California.

Hill, F., Reichart, R., and Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jurgens, D., Mohammad, S., Turney, P., and Holyoak, K. (2012). SemEval-2012 Task 2: Measuring Degrees of Relational Similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada.

Karpinska, M., Li, B., Rogers, A., and Drozd, A. (2018). Subcharacter Information in Japanese Embeddings: When Is It Worth It? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia.

Leviant, I. and Reichart, R. (2015). Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *ArXiv*, abs/1508.00106.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Yoshua Bengio et al., editors, *1st International Conference on Learning Representations, ICLR 2013,* Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia.

Nikulásdóttir, A. B., Guðnason, J., and Steingrímsson, S. (2017). *Language Technology for Icelandic. Project Plan.* Icelandic Ministry of Science, Culture and Education.

Oliveira, H. G., Sousa, T., and Alves, A. O. (2020). TALES: Test Set of Portuguese Lexical-Semantic Relations for Assessing Word Embeddings. In *HI4NLP@ECAI.*

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*, Doha, Qatar.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta.

Santus, E., Yung, F., Lenci, A., and Huang, C.-R. (2015). EVALution 1.0: an Evolving Semantic

[2] https://github.com/stofnun-arna-magnussonar/ordgreypingar_embeddings

[3] http://embeddings.arnastofnun.is/

[4] https://almannaromur.is/

Dataset for Training and Evaluation of Distributional Semantic Models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, Beijing, China.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, Miyazaki, Japan.

9. Language Resource References

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2021). *IceBATS – The Icelandic Bigger Analogy Test Set.* CLARIN-IS, http://hdl.handle.net/20.500.12537/120.

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2022a). *Word Embeddings – FastText optimized for IceBATS 22.04.* CLARIN-IS, http://hdl.handle.net/20.500.12537/211.

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2022b). *Word Embeddings – GloVe optimized for IceBATS 22.04.* CLARIN-IS, http://hdl.handle.net/20.500.12537/210.

Friðriksdóttir, Steinunn Rut and Daníelsson, Hjalti and Steingrímsson, Steinþór. (2022c). *Word Embeddings – Word2Vec optimized for IceBATS 22.04.* CLARIN-IS, http://hdl.handle.net/20.500.12537/209.

Steingrímsson, Steinþór and Barkarson, Starkaður. (2020). *Icelandic Gigaword Corpus 1 (IGC1) – version 20.05.* http://hdl.handle.net/20.500.12537/41.