EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D1.1: Datasets, benchmarks and evaluation metrics for cross-lingual word embeddings (T1.5)

Executive summary

The report presents collected and newly created resources to build and evaluate monolingual and cross-lingual text embeddings within the EMBEDDIA project, addressing nine languages: English, Slovene, Croatian, Estonian, Lithuanian, Latvian, Russian, Finnish, and Swedish. To build monolingual text embeddings, one requires large monolingual text corpora. To align monolingual embeddings and get cross-lingual embeddings, bilingual and multilingual resources (dictionaries, lexicons, parallel corpora) are required. Evaluation of text embeddings is performed in two ways: i) an intrinsic evaluation uses synthetic tasks and the evaluation measures deal only with a given embedding or cross-lingual transformation, and ii) an extrinsic evaluation uses embeddings in a downstream task, e.g., text classification or named entity recognition. For intrinsic evaluation, two novel resources were constructed, each consisting of datasets for several languages: a) novel word analogy collection, suitable for learning word analogies in morphologically rich languages (for all project languages), and b) graded word similarity data collection for English, Slovene, Croatian, and Finnish.

Partner in charge: UL

Project co-funded by the European Commission within Horizon 2020 Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D1.1**
Deliverable full title:	**Datasets, benchmarks and evaluation metrics for cross-lingual word embeddings**
Deliverable short title:	**Datasets and evaluation for embeddings**
Document identifier:	**EMBEDDIA-D11-DatasetsAndEvaluationForEmbeddings-T15-submitted**
Lead partner short name:	**UL**
Report version:	**submitted**
Report submission date:	**30/09/2019**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Marko Robnik-Šikonja (UL), Matej Ulčar (UL), Luka Krsnik (UL)**
Co-author(s):	**Matthew Purver (QMUL), Saturnino Luz (UEDIN)**
Status:	**_ draft, _ final, X submitted**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
11/06/2019	v1.0	Marko Robnik-Šikonja (UL)	Initial version.
19/06/2019	v1.1	Matej Ulčar (UL)	Describing the analogy task.
20/06/2019	v1.2	Matej Ulčar (UL)	Adding corpora information.
20/06/2019	v1.3	Luka Krsnik (UL)	Describing the NER task
19/06/2019	v1.4	Matthew Purver (QMUL)	Describing the SemEval 2020 task.
22/06/2019	v1.5	Marko Robnik-Šikonja (UL)	Integration of changes.
24/06/2019	v1.6	Matej Ulčar (UL)	Describing the new analogy dataset.
02/07/2019	v1.7	Marko Robnik-Šikonja (UL)	Describing evaluation metrics.
01/08/2019	v1.8	Mark Granroth Wilding (UH)	Internal review.
22/08/2019	v1.9	Senja Pollak (JSI)	Internal review.
28/08/2019	v1.10	Marko Robnik-Šikonja (UL)	Consolidation based on internal reviews.
10/09/2019	v1.11	Nada Lavrač (JSI)	Report quality checked and finalised.
23/09/2019	final	Marko Robnik-Šikonja (UL)	Report finalised.
30/09/2019	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

1. Introduction 4
2. Datasets 5
2.1 Monolingual corpora 5
2.2 Bilingual and multilingual resources 5
2.3 Combining dictionaries with sentence aligned corpora 7
3. Benchmarks 7
3.1 Baseline embeddings approach 8
3.2 Intrinsic evaluation benchmarks 8
3.2.1 Word analogy 9
3.2.2 Graded word similarity in context - SemEval 2020 shared task 11
3.3 Extrinsic evaluation benchmarks 13
4. Evaluation metrics 14
4.1 Metrics for intrinsic evaluation 14
4.2 Metrics for extrinsic evaluation 15
5. Conclusions and further work 16

List of abbreviations

NLP	Natural Language Processing
NER	Named Entity Recognition
WP	Work Package

1 Introduction

The EMBEDDIA project aims to improve cross-lingual transfer of language resources and trained models using text embeddings and cross-lingual text embeddings. **Word embeddings** are representations of words in numerical form, as vectors of typically several hundred dimensions. The vectors are used as an input to machine learning models; for complex language processing tasks these are typically deep neural networks. The embedding vectors are obtained from specialized learning tasks, based on neural networks, e.g., wordvec (Mikolov, Le, & Sutskever, 2013), GloVe (Pennington et al., 2014), or FastText (Bojanowski et al., 2017). For training, the embedding algorithms use large monolingual text collections (called corpora) that encode important information about word meaning as distances between vectors. In order to enable downstream machine learning on text understanding tasks, the embeddings shall preserve semantic relations between words, and this is true even across languages. Similarly to word embeddings, other text units, such as characters, sentences, or documents can be embedded into numeric space

Modern word embedding spaces exhibit similar structures across languages, even when considering distant language pairs like English and Vietnamese (Mikolov, Le, & Sutskever, 2013) This means that embeddings independently produced from monolingual text resources can be aligned (Mikolov, Le, & Sutskever, 2013), resulting in a common cross-lingual representation, called **cross-lingual embedding**, which allows for fast and effective integration of information in different languages.

Probably the best known word embeddings are produced by the word2vec method (Mikolov, Sutskever, et al., 2013) which we use as a baseline. The problem with word2vec embeddings is their failure to express polysemous words. During training of an embedding, all senses of a given word (e.g., *paper* as a material, as a newspaper, as a scientific work, and as an exam) contribute relevant information in proportion to their frequency in the training corpus. This causes the final vector to be placed somewhere in the weighted middle of all words' meanings. Consequently, rare meanings of words are poorly expressed with word2vec and the resulting vectors do not offer good semantic representations. For example, none of the 50 closest vectors of the word *paper* is related to science. The idea of **contextual embeddings** is to generate a different vector for each context a word appears in and the context is typically defined sentence-wise. To a large extent, this solves the problems with word polysemy, i.e. the context of a sentence is typically enough to disambiguate different meanings of a word for humans and so it is for the learning algorithms. In our work, we mostly use, analyze, and improve upon currently the most successful approaches to contextual text embeddings, ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018). Note that the state-of-the-art in embeddings is rapidly progressing (for example, at the time the EMBEDDIA project was conceived, the methods for training contextual embeddings did not exist). It is therefore possible, that working methods will change during the project duration.

The objectives of workpackage WP1 of the EMBEDDIA project are to advance cross-lingual and context-dependent word embeddings and test them with deep neural networks. The specific objectives of WP1 are as follows:

1. advance cross-lingual and multilingual word embeddings technology,
2. advance context-dependent and dynamic embeddings technology,
3. advance deep learning technology for morphologically rich, less-resourced languages,
4. improve interpretability of models and visualisation of results.

These objectives are followed within tasks T1.1–T1.4, while the aim of task T1.5 is to collect and prepare public as well as private resources, including datasets and benchmarks required to evaluate the developed monolingual and cross-lingual word embeddings. This report describes the results of the work performed in T1.5 in the first nine months of project duration.

EMBEDDIA works with **nine languages**: English, Slovene, Croatian, Estonian, Lithuanian, Latvian, Russian, Finnish, and Swedish. The repository of collected training and evaluation data is stored in a dedicated private cloud available to all project partners. The collected datasets mostly stem from publicly available sources or resources available for research purposes, with some exceptions which we clearly indicate. We tried to collect as many and as good resources as possible for all involved languages, but there are some intrinsic limitations that cannot be overcome within a single project. We therefore work with all the above mentioned languages but not on all tasks, subject to available resources.

This report is split into further four sections. In Section 2, we present collected monolingual, bilingual and multilingual datasets required to build monolingual and cross-lingual embeddings. In Section 3, we describe the benchmarks we use to compare and evaluate the constructed embeddings. In Section 4, we list the evaluation metrics used in intrinsic and extrinsic evaluation. We summarize the conclusions about the reported datasets, benchmarks, and evaluation methods in Section 5, where we also outline the plans for further work and emphasize the inherently incremental nature of dataset collection and benchmarking.

2 Datasets

The generation of high quality embeddings requires huge monolingual text corpora, mostly available through the EU CLARIN infrastructure[1]. These resources and their characteristics are presented in Section 2.1.

To align embeddings across languages and to produce cross-lingual embeddings, most approaches require additional information, contained in bilingual resources (dictionaries, lexicons, translation memories), multilingual resources (WordNets, Wikipedia, Wiktionaries), and multilingual parallel corpora, e.g., EU DGT-TM (Steinberger et al., 2014). These resources and their characteristics are presented in Section 2.2.

2.1 Monolingual corpora

Modern text embeddings like word2vec (Mikolov, Sutskever, et al., 2013), GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), and ULMFit (Howard & Ruder, 2018) are computed on large monolingual corpora. We succeeded in obtaining large monolingual corpora for all the EMBEDDIA languages. Many of them are publicly available, especially those acquired by web crawling (unfortunately, those are of relatively low quality due to noise involved in the crawling process). Most of web crawling corpora stem from the CoNLL 2017 Shared Task (Ginter et al., 2017), which covers 45 languages and also contains Wikipedia dumps. English Wikipedia may be of interest on its own due to its size, and possibility to link to from other languages, therefore, we include it separately. Project partners assured access also to some corpora with limited access (e.g., the case of Slovene, the Gigafida corpus (Logar Berginc & Iztok, 2011), which can be used only for research purposes due to copyright issues). We describe the main characteristics of the gathered corpora in Table 1.

2.2 Bilingual and multilingual resources

There are two types of approaches to align monolingual embeddings in a joint latent space. The first approach uses bilingual or multilingual supervision in the form of dictionaries, parallel corpora, or translation memories (human translated sentences or paragraphs). This approach is suitable for languages where such resources exist. For EMBEDDIA languages this is true for all the languages when paired with English, but not for all the pairs of languages. This is not surprising, hence we use English as a hub language in the project. The most frequently used supervised mapping approach is vecmap (Artetxe et

[1] https://www.clarin.eu/

Table 1: The collected monolingual corpora and their properties: size (in billion of tokens, asterisk (*) denotes words instead of tokens), availability (public or for research purposes only), and location (the linmks are clickable).

Language	Corpus	Size	Avail.	Location (clickable links)
Croatian	hrWaC 2.1	1.4	public	Clarin
	Riznica 0.1	0.1	public	Clarin
	CoNLL 2017	*0.6	public	Lindat/Clarin
English	Wikipedia 2018	*2.3	public	Lindat/Clarin
	CoNLL 2017	*9.4	public	Lindat/Clarin
Estonian	CoNLL 2017	0.3	public	Lindat/Clarin
Finnish	Ylilauta downloadable version	0.027	public	Kielipankki
	CoNLL 2017	1.0	public	Lindat/Clarin
Latvian	CoNLL 2017	0.3	public	Lindat/Clarin
Lithuanian	Wikipedia 2018	*0.024	public	Lindat/Clarin
	DGT-UD 1.0	0.071	public	Clarin
Russian	CoNLL 2017	3.2	public	Lindat/Clarin
Slovene	Gigafida 2.0	1.2	research	Clarin
	slWaC 2.0	1.2	research	Clarin
	CoNLL 2017	*0.5	public	Lindat/Clarin
Swedish	CoNLL 2017	2.9	public	Lindat/Clarin

al., 2016, 2018a). The second approach is completely unsupervised and implicitly builds a dictionary during the construction of cross-lingual embedding (Conneau et al., 2018; Artetxe et al., 2018b) This approach can be used for any language pair but is in principle inferior to supervised approaches for languages where the supervision information is available.

Below we list the collected bilingual and multilingual resources for the EMBEDDIA languages.

Dictionaries:

Apertium bilingual dictionaries are mostly machine learnt from parallel corpora. There are only a few pairs available for the EMBEDDIA languages (Finnish-English, Croatian-English, Croatian-Slovene)[2].

Taas bilingual dictionaries of English-X and X-English type. These dictionaries were automatically generated within the EU Taas project (Aker et al., 2014). There exist dictionaries from English to all EMBEDDIA languages (and inverse), except for Croatian and Russian[3].

Wiktionary bilingual dictionaries are extracted from wiktionaries. Some dictionaries, e.g., Croatian-Slovene, are made with triangulation via English, other pairs are direct[4].

BabelNet is a multilingual encyclopedic dictionary, with lexicographic and encyclopedic coverage of terms, and a network which connects over 16 million concepts and named entities. BabelNet covers 260 languages and links resources like WordNets, Wikipedias, Wiktionaries etc. Each Babel entry, called Babel synset, represents a given meaning and contains the synonyms which express that meaning in different languages[5].

Oxford English-Slovene dictionary. To test the importance of a high quality dictionary compared to open and less reliable dictionaries, we obtained research access to translation pairs from the high quality proprietary Oxford English-Slovene dictionary (300,000 pairs), which is an order of magnitude larger than any other dictionary above. This source will be used to measure the amount of supervision needed to achieve good cross-lingual embeddings.

[2] https://github.com/apertium/apertium-trunk

[3] http://metashare.tilde.com/repository/search/?q=taas+giza

[4] https://github.com/juditacs/wikt2dict

[5] https://babelnet.org/download

Parallel corpora:

DGT-UD is a translation memory based parallel corpus of 24 EU languages maintained by the EU Directorate-General for Translation. The corpus contains mainly the legislation of EU Acquis Communautaire[6].

Tatoeba datasets contain sentence aligned parallel corpora of relatively short and simple sentences, around 1000 per language pair. The datasets include each EMBEDDIA language paired with English, but not with each other (i.e. English is always one of the languages in any pair)[7].

OPUS is based on open subtitles which are sentence aligned. There are datasets for all EMBEDDIA languages[8].

EUR-Lex is a large, paragraph-aligned corpus of EU law texts, covering a vast area of subjects (Baisa et al., 2016). It is available in 24 official languages of the EU[9].

2.3 Combining dictionaries with sentence aligned corpora

As a work in progress, we report preparation of a dataset for alignment of contextual embeddings. Bilingual dictionaries can be directly used for alignment of non-contextual embeddings like word2vec or GloVe, but not for contextual word embeddings like ELMo, BERT, or ULMFiT. Contextual embeddings produce many different vectors for each word based on its context (i.e. a sentence it appears in). While this is an important characteristic that solves the problem of polysemous words, it also makes cross-lingual alignment much more difficult. In principle, to align two vectors produced with the contextual embedding, we must know not only that they represent matching words but also that they appear in matching sentences. Currently, the cross-lingual mappings of contextual embeddings compute the median point of all the vectors representing a single word in the two languages and then align the medians in the same way as they would align non-contextual vectors (Schuster et al., 2019). This approach assumes that shapes of word clusters in two mapped languages are identical, which is unjustified and may lead to suboptimal cross-lingual transformations. Using bilingual dictionaries in combination with sentence aligned corpora (see Section 2.2), in EMBEDDIA, we plan to construct a dataset enabling alignment of words in the context of sentences, to be used a dataset for cross-lingual mapping and its benchmarks.

3 Benchmarks

The embeddings created in the EMBEDDIA project will be compared with existing baseline embedding approaches on a selection of relevant benchmarks. We describe the baseline embeddings in Section 3.1 and the benchmarks in Sections 3.2 and 3.3.

To compare the quality of generated **monolingual and cross-lingual embeddings** we collected several benchmark datasets. There are two types of text embeddings evaluations.

1. Intrinsic evaluation uses synthetic tasks and the evaluation metrics deal only with a given embedding or a cross-lingual transformation. These type of evaluation is typically faster and can often be used to guide the construction of the embedding. Unfortunately, this evaluation may not show how well a particular embedding will perform on a certain downstream task. We present selected datasets and benchmarks for intrinsic evaluation in Section 3.2.

[6] http://hdl.handle.net/11356/1197
[7] https://tatoeba.org/eng/downloads
[8] http://opus.nlpl.eu/
[9] https://www.sketchengine.eu/eurlex-corpus/

2. Extrinsic evaluation uses embeddings as inputs to downstream text mining tasks, e.g., text classification or named entity recognition. The extrinsic evaluation tasks, mainly addressed in Section 3.3, are also part of the evaluation scenarios of the planned improvements in **neural networks, visualizations, and explanation methodology**. We expect that these improvements will be primarily benchmarked on tasks from WP2-WP5. We omit the description of relevant datasets from this report as they are described in deliverables D2.1 (Datasets, benchmarks and evaluation metrics for advanced cross-lingual NLP technology), D3.1 (Datasets, benchmarks and evaluation metrics for cross-lingual user generated content filtering and analysis), D4.1 (Datasets, benchmarks and evaluation metrics for cross-lingual content analysis), and D5.1 (Datasets, benchmarks and evaluation metrics for multilingual text generation). In this report, however, we present a particular NER (named entity recognition) benchmark, that we collected within WP1, which is described in Section 3.3.

3.1 Baseline embeddings approach

To benchmark the embeddings and cross-lingual embeddings, which will be created in the EMBEDDIA project, we first need to define the baseline embeddings. For this aim, we take standard, well-established, and widely-used pre-trained embeddings suitable for morphologically rich languages from the fastText library[10]. The fastText repository contains pre-trained word vectors for 157 languages (Grave et al., 2018). This includes all EMBEDDIA working languages, trained on web crawl and Wikipedia data using the fastText library. The fastText models (Bojanowski et al., 2017) were trained using an improved word2vec continuous bag-of-words (CBOW) algorithm (Mikolov, Sutskever, et al., 2013). In fastText, each word is represented as a bag of character n-grams, which is convenient for morphologically rich languages. A vector representation is associated to each character n-gram and word representations are computed as the sum of these representations. This allows computation of word representations also for words that did not appear in the training data. The parameters used in training of fastText embeddings are as follows: position-weights, dimension 300, character n-grams of length 5, window of size 5, and 10 negative samples.

3.2 Intrinsic evaluation benchmarks

Intrinsic evaluation of word embeddings and cross-lingual embeddings uses synthetic tasks to measure the distance between embedded vectors. The intrinsic evaluation tries to measure how well the notion of word similarity (distance) according to humans is emulated in the vector space. We are interested in multi-lingual word similarity datasets applicable to EMBEDDIA languages. We designed two such datasets, the first one is the well-known word analogy task which we extended from English to all EMBEDDIA languages (see Section 3.2.1), and the second one is a new context dependent similarity dataset (described in Section 3.2.2), which was accepted as a shared task for the SemEval 2020 competition, to be organized by the EMBEDDIA partners.

The advantage of intrinsic evaluation is that it is faster than using embeddings in downstream tasks, and can also guide the construction of embeddings. In practice, this sort of evaluation may not show how well a particular embedding will perform on a certain downstream task but shows reasonable correlations with several downstream tasks (Schnabel et al., 2015).

Cross-lingual word similarity datasets are affected by the same problems as their monolingual variants (Søgaard et al., 2019): the datasets evaluate semantic rather than task-specific similarity, they correlate only weakly with the performance on downstream tasks, and they do not account for polysemy. We plan to address the last problem by using bilingual dictionaries in combination with sentence aligned corpora

[10] https://fasttext.cc/

to construct a dataset in which we will align words in the context of sentences and thereby address polysemy, as described in Section 2.3.

3.2.1 Word analogy

The word analogy task was popularized by Mikolov, Sutskever, et al. (2013). The goal is to find a term y for a given term x so that the relationship between x and y best resembles the given relationship $a : b$. There are two main groups of categories: semantic and syntactic. To illustrate a semantic relationship, consider for example that the word pair $a : b$ is given as “Finland : Helsinki”. The task is to find the term y corresponding to the relationship “Sweden : y”, with the expected answer being $y =$ Stockholm. In syntactic categories, the two words in a pair have a common stem (in some cases even same lemma), with all the pairs in a given category having the same morphological relationship. For example, given the word pair “long : longer”, we see that we have an adjective in its base form and the same adjective in a comparative form. That task is then to find the term y corresponding to the relationship “dark : y”, with the expected answer being $y =$ darker, that is a comparative form of the adjective dark.

In the vector space, the analogy task is transformed into vector arithmetic and search for nearest neighbours, i.e. we compute the distance between vectors: d(vec(Finland), vec(Helsinki)) and search for word y which would give the closest result in distance d(vec(Sweden), vec(y)). In our dataset the analogies are already pre-specified, so we are measuring how close are the given pairs.

We composed the analogy datasets for all EMBEDDIA languages based on the English dataset by Mikolov, Chen, et al. (2013) [11]. Due to English-centered bias of this dataset, our dataset was first written in Slovene language and then translated into other languages where possible as explained below. Where a suitable translation for a certain word pair relationship could not be found, an alternative word pair was used. For example, if a given word pair in Slovene is "drag : dražji", its English translation is "expensive : more expensive". Since we are limited to single-word terms, we have to discard that translation and replace it with another one, with either a similar meaning "costly : costlier" or a completely different one, like "high : higher". Note, that due to language differences, the produced datasets are not aligned across languages. We wanted to assure consistency and allow the use of the datasets in cross-lingual analogies (described at the end of this subsection). For this reason, our datasets (even the English one) are somewhat different from the one by Mikolov, Chen, et al. (2013). We removed or edited some categories and added new ones to avoid or limit English-centric bias in the following way.

- We merged two categories dealing with countries and their capitals ("common capital cities" and "all capital cities") into one category.
- We changed "city in US state" category to "city in country" and used mostly European countries with a better chance to appear in the corpora of respective languages.
- We removed the category "currency", as only a handful of currencies are present in news and text corpora with sufficient frequency.
- We added two new semantic categories, "animals" and "city with river" described below.
- We added a syntactic category comparing noun case relationships.

The analogy tasks are composed of 15 categories: 5 semantic and 10 syntactic/morphological. The categories contained in our datasets are the following:

capitals and countries, capital cities in relation to countries, e.g., Paris : France,

family, a male family member in relation to an equivalent female member, e.g., brother : sister,

city in country, a non-capital city in relation to the country of that city, e.g., Frankfurt : Germany,

[11] `http://download.tensorflow.org/data/questions-words.txt`

animals, species/subspecies in relation to their genus/familia, following colloquial terminology and relations, not biological, e.g., salmon : fish,

city with river, a city in relation to the river flowing through it, e.g., London : Thames,

adjective to adverb, an adverb in relation to the adjective it is formed from, e.g., quiet : quietly,

opposite adjective, the morphologically derived opposite adjective in relation to the base form, e.g., just : unjust, or honest : dishonest,

comparative adjective, the comparative form of adjective in relation to the base form, e.g., long : longer,

superlative adjective, the superlative form of adjective in relation to the base form, e.g., long : longest,

verb to verbal noun, noun formed from verb in infinitive form, e.g., to sit : sitting; in Estonian and Finnish -da infinitive and first infinitive forms are used respectively; in Swedish present participle that functions as noun is used in place of verbal noun,

country to nationality of its inhabitants, e.g., Albania : Albanians,

singular to plural, singular form of a noun in relation to the plural form of the noun, e.g., computer : computers; indefinite singular and definite plural are used in Swedish,

genitive to dative, a genitive noun case in relation to the dative noun case in respective languages, e.g. in Slovene ceste : cesti, singular is used for all words, except "human" (or equivalent in other languages), which appears in both singular and plural; in Finnish and Estonian, dative has been replaced with allative case, the category is not applicable to Swedish and English,

present to past, 3rd person singular verb in present tense in relation to 3rd person singular verb in past tense, e.g., goes : went; in Slovene, Croatian and Russian the masculine gender past tense is used, in other languages the "simple" past tense/preterite is used,

present to other tense, 3rd person singular verb in present tense in relation to 3rd person singular verb in various tenses, e.g., goes : gone; the other tense in Slovene, Croatian and Russian is feminine gender past tense, in Finnish, Estonian and English it is present/past perfect participle, in Swedish it is supine, in Latvian and Lithuanian it is future tense.

The details on the number of analogies in different languages[12] is contained in Table 2. The analogy datasets will be submitted to the CLARIN repository.

Table 2: The constructed word analogy datasets for EMBEDDIA languages and their sizes in number of pairs.

Language	Size
Croatian	19416
English	18530
Estonian	18372
Finnish	19462
Latvian	20138
Lithuanian	20022
Russian	19976
Slovene	19918
Swedish	18480

Cross-lingual analogies: Cross-lingual word embeddings have two or more languages in the same semantic vector space. Cross-lingual word analogy task has been proposed by Brychcín et al. (2019) as an intrinsic evaluation of cross-lingual embeddings. Following their work, we compose cross-lingual analogy datasets, so that one pair of related words is in one language and the other pair from the same

[12] The dataset `http://download.tensorflow.org/data/questions-words.txt` has 19544 relations, but uses slightly different categories. We translated the Slovene dataset into English to keep datasets more similar across languages, especially for the use in cross-lingual analogy tasks.

category is in another language. For example, given the relationship in English father : mother, the task is to find the term y corresponding to the relationship brat (brother) : y in Slovene. The expected answer being $y =$ sestra (sister). We limited the cross-lingual analogies to the categories that both languages in any given pair have in common (e.g., not mixing the Latvian future with the Estonian participle).

3.2.2 Graded word similarity in context - SemEval 2020 shared task

Most intrinsic evaluation methods for embeddings do not take context into account, but are based only on properties of words in isolation; for example, on the ability of an embedding model to predict human judgements of similarity between pairs of words as recorded in resources like SimLex (Hill et al., 2015). Some recent work has introduced context-dependence into this kind of intrinsic evaluation, by measuring similarity between uses in different sentential contexts (Huang et al., 2012; Pilehvar & Camacho-Collados, 2018); however, so far this has assumed that the object of study for evaluation purposes is words with distinct discrete meanings (*polysemous* words), and so is not fully suitable for evaluation of embedding models that assign different representations to words in all contexts, or the ability of these models to reflect the subtle, graded changes in meaning that humans perceive. We have therefore developed a new evaluation task, designed to solve these problems and allow a full intrinsic evaluation of context-dependent embeddings in terms of word similarity. The organizers of the SemEval 2020 challenge[13] accepted our proposal for the task, named *Graded Word Similarity in Context (GWSC)*[14]; this will be run as a public competition in 2019-20, with the dataset we create being publicly released.[15]

The goal of GWSC is to predict graded word similarity in context, for multi-lingual data. Systems entered for the task will be presented with a paragraph of text, and must predict human judgements of the similarity of meaning of two words appearing within that context. Each pair of words will be presented in two different contexts, and thus paired with two corresponding different gold standard judgements; contexts will be chosen so as to encourage different perceptions of similarity, and models must therefore be context-aware in order to perform well on the task. The task will be multi-lingual, with datasets provided in four EMBEDDIA languages (English, Slovenian, Croatian, Finnish). The examples will not be restricted to polysemous words but include examples of more subtle, graded changes in meaning.

Our datasets will be based on pairs of words from SimLex-999 (Hill et al., 2015); the reliability and common use of this dataset makes it a good starting point and allows comparison of judgements and model outputs to the context independent case. The pairs will be translated into each language, substituting where necessary if sufficient data for context cannot be found. There will be four datasets, one per language. Each will consist of 333 pairs, each pair rated within two different contexts. The task is to be unsupervised. Since our interest is in graded change in meaning between the two different contexts, we are prioritising a high number of annotators per pair over a high number of pairs. This will help ensure high quality annotations and reduce the effects of noise.

We will gather human judgements of pairwise word similarity, using the same scale as SimLex. We will adapt the SimLex annotator instructions in order to benefit from its tested method of explaining how to focus on similarity rather than relatedness or association. Annotation will be crowdsourced for English; for the other languages we expect to recruit annotators directly (and this will be a backup strategy for English if better inter-annotator agreement is needed, although pilot studies suggest crowdsourcing is suitable). For each word pair and context, annotation requires two steps. First the annotators will be presented with a short paragraph of text (see the example in Figure 1). The texts (taken from Wikipedia) will be chosen to include both words in the target pair (although these will not initially be indicated to the reader). The annotators will be asked to read this paragraph and come up with two words inspired by it. These words can describe the topic, be related to the context or simply come to mind while reading it; the intention of this step is to ensure that the annotators have properly read and considered the paragraph

[13] https://www.aclweb.org/portal/content/semeval-2020-international-workshop-semantic-evaluation

[14] http://embeddia.eu/2019/07/17/shared-task-at-semeval-2020-organized-by-embeddia/

[15] See https://competitions.codalab.org/competitions/20905

(the results can also be used in data filtering – see below). The reason that the target words are not marked when reading the context paragraph is to help ensure that the the annotators read the complete paragraph, rather than focusing only on the target word pair. Having done that, the second step is to rate the similarity between the target pair of words that were contained in the paragraph. Reliability of annotations will be ensured by an adapted version of SimLex's post-processing, which includes rating calibration, checkpoint questions and the filtering of annotators with very low correlation to the average rating. In addition, we will use responses to the first annotation question to check annotator engagement with the context text and thus filter low quality raters. All data will be taken from Wikipedia to ensure that the dataset can be freely distributed.

Word1: population Word2: people	**SimLex: μ 7.68 σ 0.80**
Context1	**Context1: μ 6.49 σ 1.40**
Disease also kills off a lot of the gazelle population. There are many people and domesticated animals that come onto their land. If they pick up a disease from one of these domesticated species they may not be able to fight it off and die. Also, a big reason for the decline of this gazelle population is habitat destruction. People go out and cut down the branches of the trees that these gazelles need to feed from.	
Context2	**Context2: μ 7.73 σ 1.77**
But the discontent of the underprivileged, landless and the unemployed sections remained even after the reforms. The crumbling industries give rise to extreme unemployment, in addition to the rapidly growing population. These people mostly belong to the SC/ST or the OBC. In most cases, they join the extremist organizations, mentioned earlier, as an alternative to earn their livelihoods.	

Figure 1: Example word pair with two contexts, also showing mean and standard deviation of human similarity judgements from our pilot study, together with the SimLex equivalent values for comparison. Note that the human perception of similarity changes between the two contexts (it is higher for context 2 than for context 1), even though the target word pair remains the same.

We will evaluate performance on two subtasks and a baseline:

Predicting Ratings: participating systems must predict the absolute similarity rating for each pair in each context. This will be evaluated using Spearman correlation with gold-standard judgements, following the standard evaluation methodology for similarity datasets (Hill et al., 2015; Huang et al., 2012).

Predicting Changes: participating systems must predict the change in similarity ratings between the two contexts. This will be evaluated using two metrics: binary accuracy of predicting direction of change; and uncentered Pearson correlation to measure overall accuracy. We use the uncentered correlation to allow for differences in scaling while maintaining the effect of direction of change. On this subtask, any context-independent model will predict no change between contexts, and therefore score the same as a random baseline.

Baselines: we will provide five baselines based on cosine distances between word embeddings: standard word2vec embeddings as a context-independent model; context-dependent ELMo and BERT models on their own; and the concatenation of word2vec and ELMo/BERT embeddings.

So far, we have developed a procedure for automatically finding candidate contexts and for crowdsourcing judgements of the similarities of the target words contained within them. We used this to produce a pilot dataset in English (publicly released as the SemEval trial dataset)[16]. We are now in the process of translating the word pairs for the Slovenian and Finnish datasets (the translation for Croatian already exists (Mrkšić et al., 2017)), and pre-processing Wikipedia texts for all languages for context candidate discovery.

[16] https://competitions.codalab.org/competitions/20905

3.3 Extrinsic evaluation benchmarks

The final evaluation of embeddings and cross-lingual embeddings shall be their performance on downstream tasks provided by EMBEDDIA media partners or on similar media related publicly available datasets, e.g., on text classification and text generation tasks. Since none of these tasks will be available in all EMBEDDIA languages, we decided to test the embeddings on a popular downstream task of Named Entity Recognition (NER). We obtained labelled datasets for all EMBEDDIA languages. The details of these datasets are presented in Table 3, but further NER datasets are covered in deliverable D2.1.

The labels of the NER datasets of this deliverable are simplified to a common label set of three labels, present in all the addressed working languages. Therefore these datasets are not included in deliverable D2.1, which addresses more complex NER scenaria.

NER is an information extraction task that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. Take the following sentence as an example.

> [ORG NATO] Secretary-General [PER Jens Stoltenberg] is expected to travel to [LOC Washington, D.C.] to meet with [LOC U.S.] leaders.

This sentence contains four named entities: *NATO* is an organization, *Jens Stoltenberg* is a person, *Washington, D.C.* and *U.S.* are locations. The NER datasets for EMBEDDIA languages in Table 3 vary in the used label sets, some using more specific labels than others, like job, nicknames etc. There are only three labels, (*LOC*, *ORG*, and *PER*) in the intersection of all label sets. Due to this diversity in annotations and to make comparison sensible across languages, we decided to trim labels in all datasets down to these three classes.

Besides the label set used, there are further differences among the datasets. Although they are all extracted from media publications, some were built specifically for the NER task and contain high density of named entity terms, while others were originally meant for other tasks (like POS tagging) and later adapted for NER. This is the reason why the number of sentences in Table 3 is not a good indicator of the information content for the NER task and we also included the number of tags.

Each word in NER datasets is annotated with either named entity label or OTHR. Most but not all of the datasets are tagged in a way that enables detection of multi-word named entities, e.g., the words (*Jack Smith Parker* is tagged as *B-PER I-PER I-PER*, where *B-PER* marks the beginning of the named entity and *I-PER* marks continuation of the same named entity. Again, to assure comparison across all EMBEDDIA languages, we will use less specific tags compatible with all datasets, e.g., the words in (*Jack Smith Parker* are tagged with *PER PER PER*).

Table 3: The collected datasets for NER task and their properties: number of sentences, number of tagged words, availability, and link to the corpus location).

Language	Corpus	Sentences	Tags	Avail.	Location
Croatian	hr500k	25000	29000	public	link
English	CoNLL-2003 NER	21000	44000	public	link
Estonian	Estonian NER corpus	14000	21000	public	link
Finnish	FiNER data	14500	17000	public	link
Latvian	LV Tagger train data	10000	11500	public	link
Lithuanian[17]	TildeNER	NA	NA	limited	NA
Russian	factRuEval-2016	5000	9500	public	link
Slovene[18]	ssj500k	9500	9500	public	link
Swedish	Swedish NER	8500	7500	public	link

4 Evaluation metrics

In this section we describe the metrics commonly used to compare embeddings used in intrinsic (Section 4.1) and extrinsic tasks (Section 4.2). As a baseline we will use standard word2vec embeddings (Mikolov, Sutskever, et al., 2013) as a context-independent model. As context-dependent embeddings we will use the ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) models.

4.1 Metrics for intrinsic evaluation

Intrinsic evaluation of word embeddings and cross-lingual embeddings uses synthetic tasks to measure the distance between embedded vectors. The intrinsic evaluation tries to measure how well the notion of word similarity (distance) according to humans is emulated in the vector space.

As the distance function in the **word analogy task**, typically the cosine distance is used (which is more appropriate for high dimensional spaces compared to the Euclidean distance). We can compute the distance between two vectors, x and y, using dot product and cosine distance

$$d_{cos}(x,y) = 1 - \frac{x \cdot y}{|x||y|}.$$

In the first task of the **GWSC challenge** (predicting ratings), machine learning models will predict the similarity rating for each pair of words in two contexts. This will be evaluated using the Spearman correlation coefficient. In the second task (predicting changes), the models will predict the change in similarity ratings between the two contexts. This will be evaluated using classification accuracy and uncentered Pearson correlation coefficient.

The evaluation of **cross-lingual embeddings** shall measure the appropriateness of matching pairs between two languages. The comparison metric shall give higher score to cross-lingual mappings where the nearest neighbor of a source word, in the target language, is more likely to have as a nearest neighbor this particular source word. For example, let us assume that we have a collection of word pairs from a dictionary and we want to use them to evaluate cross-lingual word embedding. We take a pair of words, x_s in a source language and y_t in a target language and compute the cross-lingual mapping of the source word vector to the target embedding space. We search for the nearest words to that point. For **iNN measure** (e.g., 1NN, 5NN, or 10NN), we calculate the percentage of correct target words found in the i neighbourhood of the mapped point.

This measure may be problematic, as nearest neighbors are by nature asymmetric: point y being a k-NN of point x does not imply that x is a k-NN of y. For example, some vectors, called hubs, are with high probability nearest neighbors of many other points, while others (anti-hubs) are not nearest neighbors of any point. To solve the problem of k-NN asymmetry, Conneau et al. (2018) proposed a metric, called **CSLS** (Cross-domain Similarity Local Scaling). The idea is to construct a bi-partite neighborhood graph, in which each word of a given dictionary is connected to its k nearest neighbors in the other language. Let x_s be a word in the source language and W be a cross-lingual mapping matrix which transforms x_s into the target embedding space Wx_s. Let $N_T(Wx_s)$ be the neighborhood on this bi-partite graph, associated with a mapped source word embedding Wx_s (i.e. in the target space). Note that all k elements of $N_T(Wx_s)$ are words from the target language. Similarly, let y_t be the word in the target language and $N_S(y_t)$ be the neighborhood associated with a word y_t of the target language. The mean similarity of a source embedding x_s to its target neighborhood is denoted as

$$r_T(Wx_s) = \frac{1}{k} \sum_{y_t \in N_T(Wx_s)} d_{cos}(Wx_s, y_t).$$

[17] At the time of writing we have yet to obtain this corpus, therefor specified data information about it is somewhat scarce.

[18] The Slovene ssj500k originally contains more sentences, but only 9500 are annotated with NER data.

Similarly, we denote by $r_S(y_t)$ the mean similarity of a target word y_t to its neighborhood. These scores are computed for all source and target word vectors using an efficient nearest neighbors implementation, e.g., (Johnson et al., 2017). CSLS measure combines them into a similarity measure between mapped source words and target words

$$CSLS(Wx_s, y_t) = 2d_{cos}(Wx_s, y_t) - r_T(Wx_s) - r_S(y_t)$$

CSLS increases the similarity associated with isolated word vectors compared to iNN measure and decreases the similarity of vectors lying in dense areas.

4.2 Metrics for extrinsic evaluation

The extrinsic tasks we will use for the evaluation of embeddings and cross-lingual embeddings are essentially machine learning classification tasks, e.g., classification of labels in named entity recognition. Therefore, we will use standard evaluation methodology and metrics from text classification literature.

We will split the evaluation data into two sets, training set and testing set, and estimate the predictive accuracy of models on the testing set. For small datasets where a hold-out set would significantly reduce the learning capability due to wasted training data, we will use cross-validation approach.

In a binary classification problem, let E denote the set of all training instances, P denote the set of positive instances, and N the set of negative instances, where $P \cup N = E$ and $|P| + |N| = |E|$. Let $TP \in P$ (true positives) be a set of positive instances that are correctly classified by the learned model, $TN \in N$ (true negatives) be a set of correctly classified negative instances, $FP \in N$ (false positives) be a set of negative instances that are incorrectly classified as positives by the learned model, and $FN \in P$ (false negatives) be a set of positive instances incorrectly classified as negative instances.

Typical metrics used in text classification are:

Classification accuracy. Classification quality of the learned models is measured by the classification accuracy that is defined as the percentage of the total number of correctly classified examples in all classes relative to the total number of tested examples. In case of binary classification problem, the accuracy of a model is computed as

$$Accuracy = \frac{|TP| + |TN|}{|E|}$$

Note that the accuracy measures the classification accuracy of the model on both positive and negative examples of the target class of interest. Instead of accuracy, results are often presented with *classification error*, which is

$$Error(Model) = 1 - Accuracy(Model)$$

Precision, recall, and F-measure In binary classification, precision is the fraction of correctly classified positive instances among all predicted as positives, i.e.

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

while recall (also known as sensitivity) is the fraction of positive instances over the total amount of positive instances.

$$Recall = \frac{|TP|}{|TP| + |FN|} = \frac{|TP|}{|P|}$$

A measure that combines precision and recall in a harmonic mean of precision and recall is called F_1 measure or balanced F-score:

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

5 Conclusions and further work

We presented the resources collected in order to build and evaluate embeddings and cross-lingual embeddings in WP1 of the EMBEDDIA project. The final collection is large as we adapted several existing approaches to new languages and also designed new evaluation tasks. Nevertheless, we plan to further extend the collection of datasets during the course of the project. The research of embeddings and cross-lingual embeddings is rapidly progressing and new research results may require additional semantic resources and could bring better evaluation metrics. For example, we envisage that inclusion of concept ontologies might be beneficial both for the nconstruction as well as for the validation of embeddings. As word embeddings may capture different biases expressed in the training corpora (e.g., gender bias), there are several attempts to debias word embeddings (Bolukbasi et al., 2016). However, existing biases might form a possible intrinsic evaluation aspect.

References

Aker, A., Paramita, M. L., Pinnis, M., & Gaizauskas, R. (2014). Bilingual dictionaries for all EU languages. In *LREC 2014 proceedings* (pp. 2839–2845).

Artetxe, M., Labaka, G., & Agirre, E. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2289–2294).

Artetxe, M., Labaka, G., & Agirre, E. (2018a). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-second AAAI conference on artificial intelligence.*

Artetxe, M., Labaka, G., & Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 789–798).

Baisa, V., Michelfeit, J., Medveď, M., & Jakubíček, M. (2016). European Union language resources in Sketch Engine. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 2799–2803).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349–4357).

Brychcín, T., Taylor, S., & Svoboda, L. (2019). Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Systems with Applications*.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *Proceedings of international conference on learning representation (ICLR).*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ginter, F., Hajič, J., Luotolahti, J., Straka, M., & Zeman, D. (2017). *CoNLL 2017 shared task - automatically annotated raw texts and word embeddings.* Retrieved from `http://hdl.handle.net/11234/1-1989`

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (LREC 2018).*

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

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 328–339).

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1* (pp. 873–882).

Johnson, J., Douze, M., & Jégou, H. (2017). Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*.

Logar Berginc, N., & Iztok, K. (2011). Gigafida–the new corpus of modern Slovene: what is really in there. In *The second conference on Slavic corpora.*

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint 1301.3781*.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint 1309.4168*.

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

Mrkšić, N., Vulić, I., Ó Séaghdha, D., Leviant, I., Reichart, R., Gašić, M., ... Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, *5*, 309–324.

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pilehvar, M. T., & Camacho-Collados, J. (2018). WiC: 10,000 example pairs for evaluating context-sensitive representations. *arXiv preprint arXiv:1808.09121*.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 298–307).

Schuster, T., Ram, O., Barzilay, R., & Globerson, A. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Søgaard, A., Vulić, I., Ruder, S., & Faruqui, M. (2019). *Cross-lingual word embeddings* (Vol. 12) (No. 2). Morgan & Claypool Publishers.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., & Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpora. In *Language resources and evaluation* (Vol. 48, pp. 679–707). Springer.