Social Computing Group
Department of Informatics
Technical University of Munich

Training Domain Specific Multilingually Aligned Word Embeddings

Xinpeng Wang, Cheng Fan, and Maximilian Frantzen

Department of Informatics, Technical University of Munich

July 31, 2021

Abstract — In recent years a lot of different approaches, supervised and unsupervised, have been developed to efficiently train multilingual word embeddings. However, they often highly depend on the quality of the original monolingual embeddings and the domain specificity of the corpus where they are trained on. We show for different embedding techniques (Word2Vec, Fasttext and GloVe), how the quality of bilingual embeddings for German and English on our organic-food dataset can be increased by fine-tuning embeddings under the constraint on getting domain specific embeddings. We demonstrate the successful fine-tuning by observing a semantic shift of the embeddings towards the organic and food domain. Moreover, by performing topic clustering we are able to retrieve domain-specific clusters from our bilingual embeddings for the source and target language. Lastly, we can show that the unsupervised MUSE mapping is on par with the supervised approach for our fine-tuned embeddings.

1 Introduction

The goal of multilingual aligned word embedding is to find a joint embedding space for different languages where similar words appear in the same regions. In figure 1 this idea is depicted for German and English embeddings and three sample words.

The authors of [6]i w realized that even distant languages show similar structures and uses parallel vocabulary to find a linear mapping between the source and target language. Since then, a lot of other approaches have been developed to obtain cross-lingual word embeddings. However, most of them depend on bilingual supervision (e.g. word lexicons). The authors of [2] propose a new methodology called "Multilingual Unsupervised and Supervised Embeddings (MUSE) that learns a linear mapping between the source and target space without any cross-lingual labeled data.

This approach highly depends on quality of the monolingual embeddings from the respective corpora. In our case the quality of the embeddings is determined by their domain specificity. We use a dataset

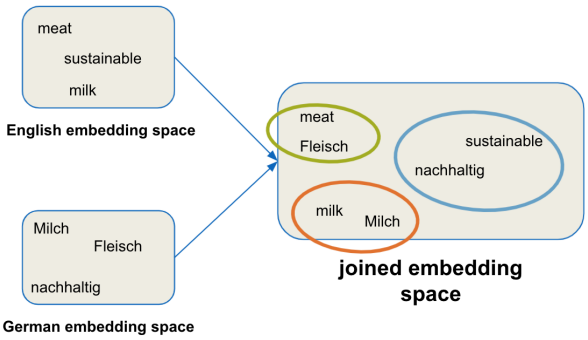

Figure 1 On the left hand side, we see the monolingual embedding spaces where similar words appear in different regions. On the right hand side, the joined embedding space for English and German maps similar words closely together.

that was curated from online articles and their respective comments on social media that deal with a range of food related topics from recipes to reports about sustainable agriculture. In order to find a high quality mapping from English to German we need to take the domain of our corpus into account. If we consider a simple example like the word "apple". If we just use pre-trained embeddings trained on data sources like Wikipedia, this could easily be misinterpreted as the company "Apple". However, the likelihood that this word appears in the context of a fruit is much higher compared with the entity "Apple" in our domain specific organic-food dataset.

Based on this assumption we tackle the following research questions:

1. How crucial is the fine-tuning of embeddings before mapping if we work with a domain specific dataset?
2. How is the embedding technique influencing the quality of the mapping?
3. How can we evaluate the quality of our domain specific fine-tuned monolingual embeddings as well as the quality of the bilingual embeddings?

In order to answer these questions we analyze different embeddings techniques for fine-tuning of the

monolingual embeddings and use MUSE as a mapping technique to obtain bilingual embeddings. For both steps, fine-tuning and mapping, we perform qualitative and quantitative evaluations to make informative decisions on the quality of the trained embeddings.

In section 2 we explain our training procedure with different embedding techniques as well as the supervised and unsupervised approaches of MUSE. Then, in section 3 we depict our experiment set-ups for training the monolingual German and English embeddings and for the linear mapping with MUSE. We provide an extensive analysis and discussion of our experiments in section 4.

Lastly, we point out open research questions and future potential in section 5.

In summary, this work contributes to the following topics:

1. We analyze the fine-tuning of GloVe, Word2Vec and Fasttext embeddings on our organic-food dataset. Thereby we evaluate the results qualitatively by looking at the most similar words of relevant words of the food domain. Moreover, we perform a quantitative evaluation by using K-means to form clusters and compute the overall coherence score.

2. We use the MUSE library and its supervised as well as its unsupervised approach as mapping technique to find a linear mapping between English and German embeddings.

3. We evaluate the mapping results by conducting topic modelling and a quantitative evaluation based on a cross-lingual similarity, word and sentence translation score.

2 Methodology

In this section, we depict the approaches on how the pre-trained embeddings were trained for Word2Vec, Glove and FastText. Moreover, we give in detail explanation of MUSE and how a linear mapping from one language to the other is computed.

2.1 Fine-tuning embeddings

Word2Vec For the Word2Vec training we utilize embeddings trained with the Wikipedia2Vec tool from [10]. The orginal embedding model was proposed by [11] and uses a conventional skip-gram model to learn, on the one hand the word embeddings and on the other hand embeddings of entities. Pre-trained Wikipedia2Vec embeddings have been applied to a wide variety of downstream tasks like text classification, question answering and named entity recognition. The pre-trained embeddings can be download here [9].

GloVe For the GloVe embeddings training we follow the implementation of [12], where we build the GloVe model with PyTorch. We save the coocurrence matrix by chunks in HDF5 binary format to avoid memory issue, and load it via a HDF5 data loader. In this way we can train the model in batch to boost the training. For English embeddings, we use the pre-trained embedding provided by the original paper [7] trained on Widipedia2014 and Gigaword5 in 300 dimensions for finetuning. For German embeddings, we use the pre-trained embeddings from open source project of Deepset [3].

FastText The pre-trained Fasttext embeddings follow the training procedure proposed in [5]. They use a mixture of Wikipedia and CommonCrawl as training data and learn 300 dimensional embeddings with a CBOW model defined by a character n-gram length of 5, a window size of 5 and negative sampling rate of 10. The authors conduct the evaluation of the trained embeddings with word analogy task and the finished pre-trained embeddings are available here [4].

2.2 MUSE

MUSE (Multilingual Unsupervised and Supervised Embeddings) is an open-source Python library for multilingual word embeddings issued by Facebook research and first published in [2]. The work presents two methods for obtaining cross-lingual word embeddings, being the supervised method and the unsupervised method respectively. Both methods aim to obtain a global linear mapping for aligning the word embeddings from the source language into the vector space of the target language. However, the supervised one requires a predefined bilingual dictionary for the training while the unsupervised one does not need any auxiliary dictionary.

2.2.1 Supervised Method

The supervised method is realised by solving the orthogonal Procrustes problem iteratively. The orthogonal Procrustes problem is an optimisation problem aiming to find a linear mapping W that maps matrix X

to matrix Y as close as possible while the mapping matrix W is restricted to be orthogonal. More concretely, it can be formulated as:

$$W^* = \underset{W}{argmin}\ ||WX - Y||_F \quad s.t.\ W^T W = I \qquad (1)$$

In our setting, the matrix X and Y represent the word embedding matrices of the source language and target language respectively. The columns of X and Y with the same index are the word embeddings in different spaces but indicate identical semantic meaning. $||WX - Y||_F$ denotes the Frobenius norm of the matrix $WX - Y$, which measures the distance between transformed word embeddings WX and target word embeddings B. $W^T W = I$ denotes the orthogonality constraint on W.

Moreover, the orthogonality constraint is introduced as an additional condition. By adding orthogonality constraint, on the one hand, the formula gets a closed-form solution by applying singular vector decomposition on YX^T. Given $YX^T = U\Sigma V^T$, it can be proved that the optimal mapping matrix is $W^* = UV^T$. On the other hand, it is stated [1] that adding orthogonality constraint boosts the performance on word translation tasks. However, the orthogonality constraint restrains the dimensionality of source embeddings and target embeddings. They must be identical in this case.

The whole algorithm performs a Procrustes solver in iterations, which can be described in the following steps:

1. Predefine a parallel dictionary of two languages (in our case English and German)

2. Build a new dictionary if it is not the first iteration

3. Obtain orthogonal mapping matrix W^* by solving the orthogonal Procrustes problem.

4. Back to step 2 until convergence

Since, the existence of a global mapping matrix between two vector spaces is assumed, the quality of the dictionary can influence the results derived from singular vector decomposition. Thereby, a new dictionary is created by each iteration. The new dictionary reserves the parallel words with high nearest neighbour performance and filters out the words pair with low nearest neighbour performance. Regarding step 4, the authors restrain the maximal iterations as convergence criteria.

2.2.2 Unsupervised Method

Adversarial Training In the unsupervised setting, the training is performed in two steps. In the first step, the authors propose to learn an initial proxy of W by using an adversarial criterion. Given $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, \ldots, y_m\}$ as two sets of n and m word embeddings from source and target language respectively, the discriminator learns to discriminate between elements randomly sampled from $W\mathcal{X} = \{Wx_1, \ldots, Wx_n\}$ and $\mathcal{Y}$. The discriminator loss can be written as:

$$\mathcal{L}_D(\theta_D \mid W) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{ source } = 1 \mid Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{ source } = 0 \mid y_i) \qquad (2)$$

where θ_D denotes the discriminator parameters.

At the same time the mapping matrix W is trained to prevent the discriminator from making accurate predictions. Thus, the loss for training W can be written as:

$$\mathcal{L}_W(W \mid \theta_D) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(\text{ source } = 0 \mid Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{ source } = 1 \mid y_i) \qquad (3)$$

Refinement Procedure After the mapping W is obtained from adversarial training, the authors of [2] report already good alignments for the embeddings. However, they propose an additional refinement step to improve the performance. Here they follow the same iterative approach as in the supervised setting. Instead of using ground truth dictionary as initialization, a synthetic parallel vocabulary is built by using the mapping W learned from adversarial training.

3 Experiments

In this section, we depict the conducted experiments that provide insights into our research questions from section 1. We present the training procedures for all three used embedding techniques: Word2Vec, GloVe and Fasttext. Furthermore, we show how the fine-tuned embeddings from the previous step are used to obtain bilingual aligned embeddings by using MUSE.

3.1 Organic-food Dataset

We have a German dataset and an English one. Both datasets contain articles and corresponding comments of users about a variety of food-related topics from different social media platforms and news outlets e.g. "chefkoch" and "zeit" for German and "nytimes" and "reddit" for English. In the pre-processing step, we performed a general text cleaning (e.g. deleting links, emojis), lemmatization of the tokens as well as stopword removal since our overall objective is to obtain domain-specific embeddings where very frequent common words would not add any useful information to the model. After the pre-processing, the German dataset has a total corpus size of 1227669 words and a vocubulary size of 192119 unique tokens. The English dataset consists of 1836072 total words and a vocabulary size of 107700 unique tokens.

3.2 Word2Vec

The overall goal of these experiments is to obtain domain specific embeddings for the English and German corpus described in the previous paragraph. These embeddings are then used in section 3.5 to obtain a mapping from the English to the German embedding space.

For training of Word2Vec embeddings we use the implementation of the gensim framework.

We perform hyperparameter optimization for the learning rate as well as for the number of epochs the model is trained. The gensim implementation uses a learning rate decay where the start learning rate and final learning rate is specified. Furthermore, we implement an early stopping mechanism based on the training loss to find the optimal number of epochs before the model starts to overfit the data. The gensim implementation also provides a parameter that ignores very rare words during the training. These words are assumed to be noise in the data and we set the threshold to a minimum count of two to not be ignored. Moreover, the model has a window size of 5 and learns the embeddings with CBOW.

We want to obtain two different kind of embeddings. First, we utilize the pre-trained Wikipedia2Vec embeddings [10] and fine-tune them on our organic-food dataset.

Secondly, we train the embeddings from scratch with randomly initialized parameters. This experiment serves as a reference for the previously trained fine-tuned embeddings.

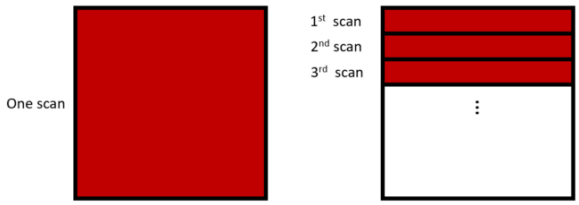

Figure 2 Instead of building up the co-occurrence matrix in single scan, we scan the matrix seperately and save each chunk to disk to save memory usage.

In all our experiments we set the embeddings dimensions to 300.

3.3 GloVe

The main issue of training the GloVe model is that building the co-occurrence matrix the computer will run out of memory, since a vocabulary of size 100,000 will lead to 10^{10} distinct co-occurence pairs. To counter this issue, we can count co-occurring pairs in multiple scans, and after each scan we save the counts to disk. In this way we can decrease the memory use while building up the co-occurrence matrix. Figure 2 shows the idea of multiple scan approach for building the co-occurrence matix. Since our co-ocurrence matrix is a large sparse matrix, we utilize the sparse gradient method provided by PyTorch, which results in a 8 times faster training time in our case.

3.4 Fasttext

For the training of Fasttext embeddings we make use of its unsupervised method to fine-tuning the embeddings. Moreover, as in the Word2Vec training we employ the gensim framework to conduct all our experiments.

For fine-tuning we download the pre-trained Fasttext embeddings from Facebook. Since the gensim implementation does not provide a training loss during training we don't have a metric during training to measure the performance of the model. This is why we need to train for different number of epochs and evaluate the results afterwards. As reference to the fine-tuning we also train Fasttext embeddings from scratch with randomly initialized parameters.

3.5 Learning linear mapping with MUSE

In order to obtain bilingual embeddings we leverage the MUSE library and its supervised and unsupervised

methodology as described in section 2.2. Regarding the network architecture we use the default size with 2048 hidden neurons and two discriminator layers. For the supervised approach we perform hyperparameter optimization regarding the refinements steps. In the unsupervised setting we additionally optimize for the number of epochs trained. We perform the experiments for the Glove and Word2Vec embeddings from the previously explained experiments in section 3.3 and 3.2 in order to compare the influence of the embedding technique on the final bilingual embeddings.

3.6 Evaluation

3.6.1 Fine-tuned Embeddings

For all three embedding techniques we conduct the same evaluation that consists of qualitative and quantitative part. In the qualitative evaluation, we look at the top most similar words based on their cosine similarity. Here, all embeddings are normalized beforehand so each vector has the same length and information of context occurrence is not taking into account. We specifically focus on domain specific words from the food and organic domain like "apple" or "sustainable" in order to check if we produce domain-specific embeddings in our training.

However, it is very difficult to make any kind of informative decision on which embedding is performing best only based on a very subjective quantitative evaluation. This is why we also conduct a quantitative evaluation based on the idea of topic coherence as described in [8]. First, we cluster the embeddings by using the K-means algorithm where each cluster can be interpreted as a topic. Then, for the top N words in each cluster its co-occurrence with all other words in the cluster is counted and divided by the overall occurrence of the target word in the document. This can be summarized in the following formula,

$$C = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}. \quad (4)$$

The intuition behind the coherence is that the higher the score the more words in each cluster are semantically related to each other. We employ the coherence score as a metric to make a quantitative decision for our trained embeddings. The gensim library already provides a method to calculate the coherence score based on formula 4 that we utitlize.

3.6.2 Quantitative evaluation for mapping

In order to evaluate the mapping quality of the different mapping methods and the embeddings, we follow the implementation of various evaluation tasks to demonstrate the effectiveness of different approaches of [2].

Word Translation This task considers the problem of retrieving the tranlation of given source words. We use the dictionary provided by MUSE library for computing the translation accuracy. For each language pair, we consider 1,500 query source and 200k target words and measure how many times one of the correct translations of a source word is retrieved, and repoart precision@k for k = 1, 5, 10.

Cross-lingual semantic word similarity We also evaluate the quality of cross-lingual word embeddings space using word similarity tasks. This task aims at evaluating how well the cosine similarity between two words of different languages correlates with a human-labeled score. We use the SemEval 2017 competition data and report the Pearson correlation.

Sentence translation retrieval We use the idf-weighted average to merge word into sentence embeddings and perform sentence retrieval on the Europarl corpus. We consider 2,000 source sentence queries and 200k target sentences for each language pair and report the precision@k for k=1, 5, 10, which accounts for the fraction of pairs for which the correct translation of the source words is in the k-th nearest neighbors.

3.6.3 Qualitative evaluation for mapping

To provide more insight about the quality of our mappings, we conduct topic modelling by K-means as qualitative evaluation, we conduct topic modelling by K-means. We perform it in following steps:

1. Conduct extra stop words removal based on the frequency of words
2. Cluster embeddings in the joint vector space into k clusters by K-Means
3. For each cluster, retrieve top 10 English and German words by measuring cosine similarity to the centroid of the cluster
4. Assign topics to each cluster depending on top words manually

To capture more domain-related clusters by K-means, we conduct an extra stop words removal on our dataset. By further analysis of the dataset, we find the majority of the organic food-related words have occurrence frequency between 500 and 15000 times regardless of the language. Thereby, for the clustering task, we only use the words in that frequency interval.

Fig.3 illustrate the loss curves obtained during supervised and unsupervised mapping for the embeddings fine-tuned by Word2Vec and GloVe respectively, where x-axis representes the number of clusters and y-axis is average cosine score for all word embeddings. The cosine score for a specific word embedding is computed as by $1 - s$ where s is the cosine similarity between the embedding and the centroid of its corresponding cluster.

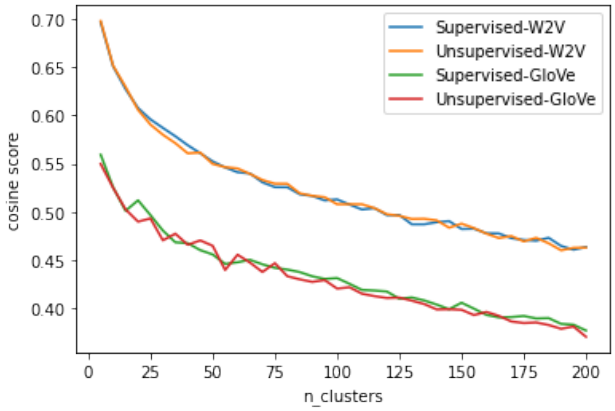

Figure 3 Loss curves obtained when tuning cluster number

By interpreting the loss curves manually, we find the elbow point located between $k = 25$ and $k = 50$ approximately. Therefore, in our experiments, we choose the cluster number $k = 35$. Consequently, we obtained 35 clusters while each of them indicates a semantic concept.

4 Results

In this section we discuss the results of the experiments described in the previous paragraphs. First, we analyse the evaluation of the fine-tuning of our embeddings. Then, we show the comparison of the different mappings based on the monolingual Word2Vec and Glove embeddings.

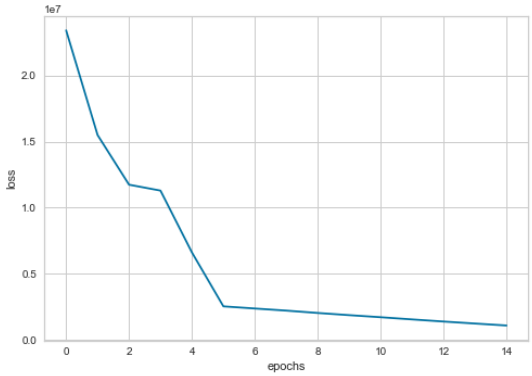

Figure 4 The loss curve of the from scratch trained Word2Vec embeddings over 15 epochs. We can observe that from epoch 5 on the loss is not significantly decreasing anymore.

4.1 Fine-tuning of monolingual embeddings

4.1.1 Word2Vec

We start with the quantitative evaluation as depicted in section 3.6 for the Word2Vec embeddings. In figure 5 the coherence scores for different hyperparameter settings for the from scratch training on the English corpus are shown. The default learning rates computed by gensim (start learning rate: 0.025 and final learning rate: 0.0001) result in the best performance when trained over nine epochs. Figure 4 illustrates the loss function of the training where we can see that the loss is not significantly decreasing from epoch six on. Our early stopping mechanism has a patience of three and thus the training was stopped at epoch nine.

In figure 6 we can see the results for the best performing English embeddings. Based on the coherence score from scratch training slightly outperforms the fine-tuning. At first sight this result seems to be very surprising since the pre-trained Wikipedia2Vec embedding was trained on a much larger corpus compared to ours. However, one has to consider that the fine-tuned embeddings are only trained for one epoch compared with nine epochs for the from scratch ones.

The coherence scores for the best performing German embeddings are shown in figure 7. Here, the optimization of the learning rate has a bigger impact than for the English embeddings. From figure 7 we can derive that the fine-tuned embedding trained over one epoch is outperforming all other embeddings.

In a next step we analyse the results from a qualitative point of view. Table 1 depicts the top ten nearest words for the word "apple" for the pre-trained Wikipedia2Vec embeddings and the best performing

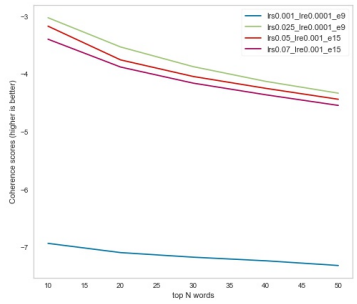

Figure 5 Coherence score for the from scratch trained Word2Vec embeddings for the top 10,20,30,40 and 50 words in each cluster. "lrs" is the start learning rate and "lre" is the final learning rate.

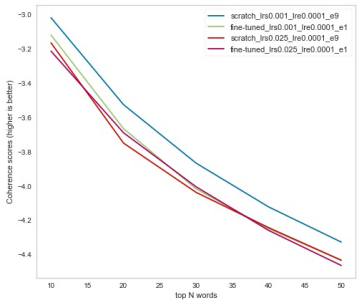

Figure 6 Coherence score for the best from scratch and fine-tuned trained Word2Vec embeddings for the top 10,20,30,40 and 50 words in each cluster. "lrs" is the start learning rate and "lre" is the final learning rate.

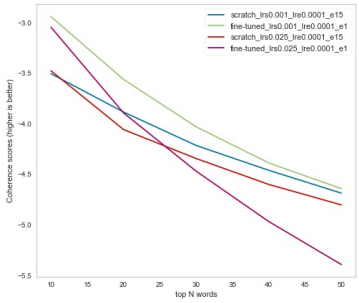

Figure 7 The best performing embeddings from the Word2Vec training for the German embeddings: The fine-tuned embedding trained over one epoch with a start learning rate of 0.001 and a final learning rate of 0.0001 outperforms all other embeddings.

from scratch trained and fine-tuned embeddings based on the quantitative analysis. The top words of the pre-trained embeddings follow a very generic terminology with very little relatedness to the food sector. At position four we can see the entity "Apple" describing the software company which is expected since the Wikipedia2Vec embeddings are trained with an entity recognition as downstream task. However, in the top words of the embeddings trained on the organic-food dataset we find exclusively tokens related to fruits and vegetables except for some noise words in the from scratch trained embedding (e.g. fuji, braeburn).

	finetuned	scratch	pretrained
0	peach, 0.56	banana, 0.57	silentype, 0.66
1	strawberry, 0.53	tomato, 0.55	paulared, 0.65
2	apricot, 0.52	strawberry, 0.54	iigs, 0.63
3	plum, 0.51	carrot, 0.54	zestar, 0.63
4	avocado, 0.51	honeycrisp, 0.53	ENTITY/Apple_Inc., 0.63
5	pear, 0.50	pear, 0.52	minnewasheta, 0.63
6	banana, 0.49	braeburn, 0.52	blackberry, 0.62
7	blueberry, 0.48	peach, 0.51	trueimage, 0.60
8	carrot, 0.48	unripe, 0.51	idared, 0.60
9	tomato, 0.47	fuji, 0.50	macterminal, 0.60

Table 1 Top 10 nearest words for the word "apple" for fine-tuned, from scratch trained and pre-trained Wikipedia2Vec embeddings. The tokens in the pre-trained column follow a very generic terminology while the ones for the fine-tuned and scratch column are semantically food-related.

Table 2 illustrates the results for word "nachhaltig" of the German embeddings. We can see the same behavior compared to the English embeddings where the top words move to a more organic context in the fine-tuned embeddings. This semantic shift from a very generic to a more food/organic-related terminology demonstrates the effectiveness of our fine-tuning on the organic-food dataset.

Thus, we can conclude that the goal of obtaining domain-specific embeddings on our organic-food dataset with Word2Vec was successful.

	finetuned	scratch	pretrained
0	umweltfreundlich , 0.52	ökologisch, 0.53	entscheidend, 0.70
1	ökologisch, 0.52	umweltfreundlich, 0.52	maßgeblich, 0.63
2	nachhaltige, 0.52	ressourcenschonende, 0.50	beeinflusste, 0.61
3	umweltschonend, 0.51	umweltverträgliche, 0.50	prägend, 0.60
4	umweltschonende, 0.50	umweltschonend, 0.47	stark, 0.58
5	umweltverträglich, 0.49	umweltgerechte, 0.47	prägten, 0.58
6	nachhaltiger, 0.49	naturverträglicher, 0.47	prägte, 0.58
7	klimafreundliche, 0.48	nachhaltigkeit, 0.47)	beeinflusst, 0.58
8	zukunftsfähige, 0.48	nachhaltige, 0.45	beeinflussend, 0.58
9	umweltgerechten, 0.47	umweltverträglich, 0.44	erheblich, 0.57

Table 2 Top 10 nearest words for the word "nachhaltig" for fine-tuned, from scratch trained and pre-trained Wikipedia2Vec embeddings.

4.1.2 GloVe

As for the evaluation procedure used for Word2Vec, we also calculate the coherence score to determine the best GloVe embeddings. We evaluate the pre-trained and fine-tuned embeddings trained for different epochs. The result is shown in figure 8 for English and German embeddings. All fine-tuned embeddings achieve a higher score than the embeddings trained from scratch. Among the fine-tuned embeddings trained from different epochs, we observe an overall better performance of the fine-tuned embeddings trained for one epoch. This is as expected since training for too long will lead to overfit on the dataset which harms the semantic meaning of pre-trained em-

beddings. Therefore we decide to choose a fine-tuned GloVe embeddings trained for one epoch for mapping.

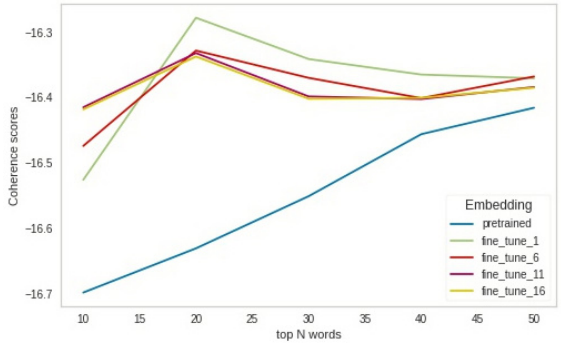

Figure 8 Coherence score for pre-trained and fine-tuned embeddings. The number of the label indicates the number of epochs the embedding are trained for.

We also train the GloVe embeddings from scratch on our organic-food dataset for reference, while the quality of the trained-from-scratch embeddings is not as expected based on the most similar words evaluation. One of the reason could be that training on GloVe model is very slow, given the limited hardware and time the model cannot learn meaningful enough word embeddings. Thus, we only use pre-trained and fine-tuned GloVe embeddings for further experiments.

Table 3 and 4 shows the qualitative evaluation of the pre-trained and the fine-tuned GloVe embeddings based on most similar words. Similar to the observation made in the case of Word2Vec, we also see that the word embeddings after fine-tuning are more related to the organic-food topic. For example, the German word "nachhaltig" is more related to organic food and environmental topics in the fine-tuned embeddings, while it has more generic meaning in the pre-trained embeddings.

	pretrained	fine-tuned
1	iphone, 0.60	pear, 0.63
2	macintosh, 0.58	banana, 0.59
3	ipod, 0.57	juice, 0.53
4	microsoft, 0.56	peach, 0.53
5	ipad, 0.56	cherry, 0.51
6	intel, 0.54	cider, 0.50
7	ibm, 0.53	orchard, 0.50
8	google, 0.53	fruit, 0.50
9	software, 0.49	carrot, 0.49
10	motorola, 0.47	sweet, 0.49

Table 3 Top 10 most similar words of word "apple" from English pre-trained and fine-tuned GloVe embeddings.

	pretrained	fine-tuned
1	entscheidend, 0.64	landwirtschaft, 0.70
2	fördern, 0.61	ökologisch, 0.64
3	maßgeblich, 0.61	produkte, 0.62
4	beeinflussen, 0.59	produktion, 0.61
5	beeinflusste, 0.58	umweltfreundlich, 0.60
6	beitragen, 0.58	wirtschaften, 0.57
7	prägend, 0.57	produkten, 0.55
8	verbessern, 0.57	ernährung, 0.55
9	gestärkt, 0.57	nutzen, 0.55
10	geschädigt, 0.56	produzieren, 0.55

Table 4 Top 10 most similar words of word "nachhaltig" from German pre-trained and fine-tuned GloVe embeddings.

4.1.3 Fasttext

The main difference between Fasttext and the other two embedding approaches is its capability to train on sub-character level. This allows Fasttext models to calculate word representations that did not appear in the training corpus. However, in our case the ability to fall back to the sub-word level leads to embeddings that are syntactically very similar but not semantically. Table 5 depicts the top 10 nearest words for "sustainable" for a fine-tuned model trained over one epoch and a window size of five. On the on hand, we can clearly see that the pre-trained model seems to capture the semantic meaning of the word. On the other hand, the closest words in fine-tuned model all contain the same character n-grams like "sustain" or "tainable". From this findings we conclude that the unsupervised fine-tuning of Fasttext embeddings on our domain specific organic-food dataset does not result in useful embeddings. This is why we did not further investigate either the results of the fine-tuning nor the mapping with MUSE with Fasttext embeddings.

	finetuned	pretrained	scratch
0	selfsustainable, 0.97	sustainably, 0.81	organicsustainable, 0.91
1	nonsustainable, 0.97	sustainability, 0.80	sustainablefood, 0.88
2	unsustainable, 0.95	renewable, 0.73	selfsustainable, 0.87
3	sustainablysourced, 0.95	compostable, 0.64	sustainablysourced, 0.83
4	sustainablefood, 0.95	ecological, 0.63	attainable, 0.82
5	sustainably, 0.95	environment, 0.63	nonsustainable, 0.81
6	organicsustainable, 0.95	unsustainable, 0.63	sustainably, 0.80
7	sustained, 0.93	ecologically, 0.63	obtainable, 0.79
8	obtainable, 0.92	ecotourism, 0.62	selfsustainability, 0.78
9	attainable, 0.91	aquaculture, 0.62	sustainabilityoriented, 0.77

Table 5 Top 10 nearest words for Fasttext embeddings: Fine-tuned over one epoch, pre-trained embeddings from Facebook and from scratch trained embeddings over 10 epochs. The two embeddings trained on our corpus only show syntactical similarity but fail to reflect the domain-specificity of our organic-food dataset.

4.2 Bilingual embeddings

In this section we discuss the results of the different mapping experiments with MUSE. In order to give an intuitive explanation on how the embeddings from difference language are mapped into the same embedding space, we show the mapping process of unsupervised training in Fig 9 visualized by the PCA method. The words of the same meaning are gradually mapped together along the training.

4.2.1 Quantitative evaluation

Table 7 shows the cross-lingual similarity score on GloVe and Word2Vec embeddings with supervised and unsupervised method. Table 8 and 9 show the word and sentence translation task respectively. Detailed discussion will be done in the comparison section.

	GloVe	Word2Vec
Supervised	0.66	0.64
Unsupervised	0.66	0.67

Table 7 Cross-lingual wordsim task, we report Pearson correlation.

	Unsupervised		Supervised	
	GloVe	Word2Vec	GloVe	Word2Vec
P@1	51.0	32.0	52.4	33.4
P@5	66.5	48.0	70.1	55.1
P@10	72.6	55.0	77.5	62.4

Table 8 English-German word translation retrieval. We report the average precision@k from 1,5k source word queries using 200k target words.

	Unsupervised		Supervised	
	GloVe	Word2Vec	GloVe	Word2Vec
P@1	53.7	35.3	57.7	41.2
P@5	73.3	54.5	77.3	61.6
P@10	79.1	60.6	82.9	68.4

Table 9 English-German sentence translation retrieval. We report the average P@k from 2,000 source queries using 200,000 target sentences.

4.2.2 Qualitative evaluation

By performing K-Means on the aligned word embeddings, we obtain 35 clusters. We categorise the clusters into three types, being domain-related clusters, meaningful clusters and garbage clusters. Domain-related clusters represent some meaningful domain aspects in our case organic food related clusters, while meaningful clusters indicate a reasonable concept but are not relevant to the organic food domain. Regarding garbage clusters, we can not assign any topics to them simply by looking at the top words they contain. If we analyze Table 10 for an example, which is a snippet of Table 12, the "Nutrition" cluster,an organic-food related cluster, contains top words like carbohydrate, sugar, protein or fiber in English and eiweiß, zucker, substanzen or kohlenhydrate in German. The top words mentioned before in the "Nutrition" cluster match the concept perfectly. Meanwhile, we regard the "Month" cluster as a meaningful cluster since it is a general concept in daily life and the German and English words fit the topic description. Moreover, table 10 presents two garbage clusters as showcases. Even both of them are assigned as garbage clusters, they are different intrinsically. The former, which contains English words put, pull, turn, bring, etc. and German words schieben, kippen, packen, holen, ziehen, etc., indicates words translation inside the cluster, whereas the latter contains arbitrary words in English and German.

Topics	Top words for English and German
Nutrition	carbohydrate sugar protein fiber nutrient eiweiß zucker substanzen kohlenhydrate
Month	october sept july june december november september oktober januar april
Garbage	put pull turn bring throw schieben kippen packen holen ziehen
Garbage	tell understand believe realize remind erwarten glauben wundern beschweren

Table 10 Part of clusters obtained by supervised method for Word2Vec Embeddings

Table 11, 12, 13 and 14 in appendix 5 present all clusters and their corresponding manually assigned topics for the final aligned word embeddings derived by supervised and unsupervised mapping methods from Word2Vec and GloVe respectively. Meanwhile, topics related to organic food are bold. Refer to them for more details.

4.2.3 Comparison of supervised and unsupervised methods

In quantitative evaluation, both supervised and unsupervised methods show similar performance on cross-lingual similarity regardless of the embedding models as Table 7 shows. Regarding the word and sentence

Mapping method	Embedding model	Domain-related clusters	Other meaningful clusters	Garbage clusters
Supervised	Word2Vec	4	19	12
Unsupervised	Word2Vec	4	17	14
Supervised	GloVe	5	16	14
Unsupervised	GloVe	5	17	13

Table 6 Number of different types of clusters

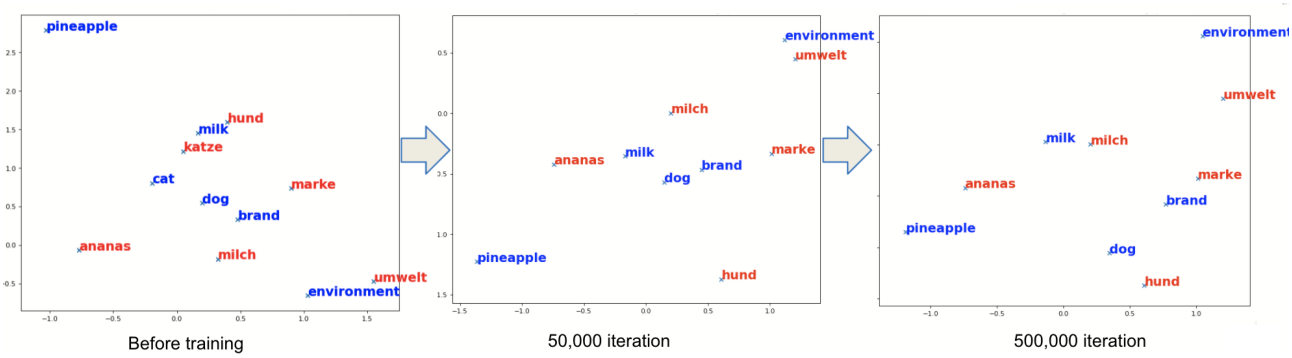

Figure 9 Visualiation of mapped source and target language embeddings at different training stage.

translation retrieval tasks, the supervised method outperforms slightly the unsupervised one for both embedding models as illustrate in table 8 and 9.

For the qualitative evaluation, by observing the number of clusters in each category in Fig. 6, we notice they are almost identical. Meanwhile, the concepts we assign for the clusters express a high similarity for both methods. With supervised mapping for Word2Vec embeddings, we obtain four organic food-related clusters, being nutrition, food, livestock, farming and with an unsupervised setting we get four clusters naming food, fruit, nutrition, farming, which are related to organic food. In this case, three out of four concepts are the same for both methods. The same observation can be made for GloVe embeddings as well. All facts mentioned above indicate a similar performance of supervised and unsupervised methods.

4.2.4 Comparison of Word2Vec and GloVe

Now we compare the performance of mapping for Word2Vec embeddings and GloVe embeddings quantitatively. The result shows that GloVe embeddings outperform Word2Vec embeddings in the cross-lingual similarity task as well as word and sentence translation retrieval tasks. Therefore, from a quantitative perspective GloVe embeddings outperform Word2Vec embeddings.

However, for the qualitative analysis the cluster number of different categories doesn't show any difference between the two embedding techniques. Both of them produce 4 or 5 domain-related clusters, 17 meaningful clusters as well as 13 garbage clusters. Additionally, by checking the top words listed in Table 12 and 14, we ensure the clustering quality of Word2Vec embeddings is as good as GloVe embeddings.

In summary, we can obtain desired clusters both with Word2Vec embeddings and GloVe embeddings, even if the GloVe embeddings show better performance on the quantitative evaluation.

5 Conclusion and Future work

In this work, we show how to obtain domain-specific bilingual embeddings for English and German. First, we fine-tune monolingual embeddings for Word2Vec and Glove and demonstrate how the pre-trained embeddings are semantically shifted to the food domain based on our organic-food dataset. Furthermore, we validate our findings with a qualitative and quantitative evaluation in order to make an informative decision on the quality of our results. These fine-tuned embeddings are then the key ingredient for the success of the linear mapping from the English embedding space into the German one. In an extensive study, we showcase that the unsupervised MUSE approach is on par with the supervised method by performing a topic clustering. Additionally, this finding is supported by different downstream tasks like cross-lingual similarity or sentence translation that serve as a quantitative evaluation of the mapping result. Finally, we can also conclude that there are no significant differences between the Word2Vec and Glove embedding techniques.

However, we were not able to fine-tune any meaningful Fasttext embeddings with its unsupervised method on our dataset. The issue of falling back to syntactically similarity despite semantically similarity remains an open research question for future work.

References

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2017.

[2] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[3] Deepset. Dockerized training of glove embeddings. https://gitlab.com/deepset-ai/open-source/glove-embeddings-de, 2019.

[4] Facebook. Library for efficient text classification and representation learning. https://fasttext.cc/docs/en/crawl-vectors.html, 2016.

[5] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[6] Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[8] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA, 2015. Association for Computing Machinery.

[9] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. https://wikipedia2vec.github.io/wikipedia2vec/pretrained/.

[10] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics, 2020.

[11] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259. Association for Computational Linguistics, 2016.

[12] Peng Yan. A comprehensive python implementation of glove. https://github.com/pengyan510/nlp-paper-implementation, 2020.

Appendix

Table 11 Clusters from supervised method for Word2Vec Embeddings

Topics	Top 10 words for English and German
Nutrition	nutrient carbohydrate toxin liquid fiber protein mineral nitrogen body sugar eiweiß nährstoffe substanzen schadstoffe kohlenhydrate körper zucker stoff gifte stoffe
Food	pickle roasted sauce basil salad sausage carrot pasta dessert chili nudeln kartoffeln karotten salat müsli bohnen früchte eis gewürze tomaten
Livestock	cattle plant livestock pasture bird cow crop grass animal manure wiese kühe weiden huhn rinder nutztiere rind acker pflanzen feld
Farming	chemical pesticide synthetic additive insecticide herbicide toxin substance gmo toxic pestiziden pflanzenschutzmittel pestizide chemikalien substanzen chemisch gifte zusatzstoffe kunstdünger gift
Month	october sept july june december september april august november january november oktober september januar april februar juni dezember märz august
Trading	sell purchase provide donate pay distribute consume afford save deliver verkaufen produzieren bezahlen ausgeben liefern anbieten importieren zahlen verbrauchen versorgen
Unit	roughly pound million half percent gallon six nine eight ten circa tonnen millionen viertel liter mio quadratmeter prozent drittel kilogramm
Shop	store supermarket retailer grocer walmart aldi safeway offering costco brand discounter supermarkt läden anbieter bioladen bioläden kunde händler supermärkte lieferanten
Trend	increase decrease reduce lower improve reduction decline low boost vary steigern erhöhen senken verringern sinken reduzieren ansteigen steigen steigend gering
Order	last next past later first ahead early ago back within nächst kommend letzt vergangen letzte spät zweimal erste innerhalb voraus
Resources	article post blog column comment page website video headline quote artikel text kommentar bericht interview posting informationen beitrag sendung bücher
Contrainer	window roof patio foot mouth jar container cart leg door tür glas mund kühlschrank augen haut nase schale flasche tüten
Amount	amount percentage number quantity concentration component portion presence level intake anteil mengen menge faktor summe anzahl verbrauch bedeutung zahl gehalt
Large Quantity	various numerous multiple largely several certain often primarily typically mainly zahlreiche divers zahlreich sämtliche hauptsächlich häufig überwiegen verschieden keinerlei meist
Argument	argument concept belief idea perspective attitude situation position notion understanding argumentation argumente einstellung argument verständnis ansatz meinung ideologie behauptungen logik
Emotion	funny happy sad stupid nice fun crazy love lucky okay lustig schlau nett tollen schön blöd traurig komisch naiv lecker
Job	director david james robert professor founder peter smith john editor geschäftsführer michael leiter martin klaus wolfgang josef chef hans präsident
Policy	regulation agency guideline usda commission certification authority fda inspection approval vorschriften richtlinien gesetze maßnahmen kontrollen gesetz vorgaben bundesregierung kommission zulassung
Family Role	child parent kid daughter son friend wife dad mother husband sohn eltern kind tochter mutter freundin mann schüler kinder vater
Illness	illness disease symptom chronic problem effect infection damage risk consequence krankheit krankheiten allergien krebs belastung auswirkungen schäden risiken auslösen ursache
Business	revenue investment profit business government subsidy taxis input competition corporation investitionen förderungen gewinn subventionen konzerne steuerzahler arbeitnehmer banken förderung gewinne
Politics	republican gop democrat democrats candidate congress politician voter conservative trump spd partei politiker regierung cdu övp fdp wähler grüne csu
Location	town village south northern region city area downtown mexico west stadt norden dorf gegend osten westen städten region südamerika berlin
Garbage	put pull turn bring throw blow toss hang walk move schieben kippen packen holen ziehen fallen werfen wandern drücken legen
Garbage	anything certainly something simply always really actually obviously definitely either wirklich irgendwie tatsächlich niemand überhaupt nie trotzdem sowas durchaus vielleicht
Garbage	already still therefore certainly however actually although simply unless yet trotzdem dennoch deshalb außerdem seitdem schließlich ohnehin deswegen letztlich somit
Garbage	oppose refer consider seem appear recognize appeal embrace defend apply bezeichnen verteidigen darstellen empfinden äußern unterstellen scheinen erscheinen betrachten formulieren
Garbage	fresh meat healthy healthful glutenfree nonorganic nutritious specialty cheap nongmo hochwertig milchprodukte hochwertige bioqualität regional biofleisch preiswert gesund bioprodukte billig
Garbage	tell understand believe realize remind worry ask argue assume talk erwarten glauben wundern beschweren merken erfahren wünschen begreifen aufregen behaupten
Garbage	sustain destroy create adapt survive implement eliminate protect convert develop stoppen fördern verhindern unterstützen bekämpfen etablieren stärken entwickeln finanzieren entziehen
Garbage	difficult important effective dangerous helpful relevant necessary easy safe useful fragwürdig sinnvoll wichtig problematisch schwierig bedenklich gefährlich unsinnig interessant umweltfreundlich
Garbage	sustainability cultural economic ecological agricultural conservation environmental social planning financial gesellschaftlich soziale umweltschutz nachhaltigkeit wirtschaftlich sozial politische nachhaltig politisch ökonomisch
Garbage	large huge small relatively vast massive significant broad strong limited erheblich enorm gewaltig riesig gering riesige extrem unglaublich grosse starke
Garbage	recognize demonstrate determine confirm conclude cite identify acknowledge define indicate berücksichtigen nachweisen festlegen beurteilen klären mitteilen überprüfen beweisen prüfen bestätigen
Garbage	society ecosystem economy environment life culture population planet diversity world wohlstand armut gesellschaft kapitalismus mensch menschheit natur lebensweise welt freiheit

Table 12 Clusters from unsupervised method for Word2Vec Embeddings

Topics	Top 10 words for English and German
Food	sausage pasta dessert sandwich roasted sauce salad steak bacon homemade nudeln müsli pizza servieren lecker kuchen kartoffeln eis salat frühstück
Fruit	basil cucumber strawberry carrot peach tomato blueberry avocado citrus lemon karotten paprika kartoffeln früchte frucht tomaten salat bohnen bananen äpfel
Nutrition	carbohydrate sugar protein fiber nutrient sodium additive preservative mineral antioxidant eiweiß zucker substanzen kohlenhydrate fettsäuren vitamine inhaltsstoffe nährstoffe stoff zusatzstoffe
Farming	crop chemical pesticide fertilizer insecticide plant herbicide livestock substance pest pestiziden pflanze pflanzenschutzmittel monokulturen organismen gifte pestizide dünger kunstdünger chemikalien
Month	october sept july june december september april november august january november september oktober januar april juni februar dezember märz august
Production	consume kill survive sell destroy contaminate absorb feed ingest convert vernichten produzieren verfüttern importieren entsorgen abbauen verarbeiten transportieren verzehren anbauen
Amount	percent percentage cost amount roughly cent million income billion revenue verbrauch summe prozent anteil gehalt circa eur milliarden quadratmeter ertrag
Container	window container roof dirt pan thick pile mouth layer oven schale haut glas mund wasser kühlschrank düsen nase flasche luft
Order	later last next early first slowly back past soon finally nächst spät kommend letzt letzte irgendwann danach langsam erste schnellen
Shop	retailer grocer store company walmart supermarket supplier customer competitor brand discounter händler anbieter lieferanten kunde supermärkte bioläden erzeuger firmen supermarkt
Resource	article post comment blog column page website headline quote video artikel kommentar text bericht posting interview beitrag informationen kommentare sendung
Argument	concept situation argument understanding position perspective belief attitude idea goal argumentation einstellung argumente verständnis ansatz widerspruch argument meinung konzept erkenntnis
Number	three six four five eight two ten several seven nine drei sechs fünf vier zwölf zwei sieben zehn neun acht
Job & Name	director robert david james professor peter founder john smith chief geschäftsführer michael klaus wolfgang leiter martin chef josef hans präsident
Trend	increase decrease reduce reduction large massive low significant huge high zunehmend steigend gering erhöhen enorm sinken verringern wachsend reduzieren steigern
Emotion	stupid sad bad ignorant funny sick happy crazy silly weird lustig schlau blöd traurig dumm naiv komisch schlimm nett seltsam
Family Role	child kid parent daughter son friend wife woman dad mother eltern kind sohn kinder mann tochter freundin mutter familie vater
Location	town downtown neighborhood yard patio road village apartment park parking straße dorf stadt wiese wohnung bauernhof mitten hofladen hof gegend
Policy	agency regulation government commission congress legislation funding federal policy usda gesetze bundesregierung maßnahmen vorschriften kontrollen gesetz förderungen richtlinien kommission regierung
Illnesss	illness disease symptom chronic effect problem infection obesity damage stress krankheit krankheiten allergien auswirkungen krebs belastung risiken schäden probleme ursache
Region	europe france canada mexico italy region india usa northern africa südamerika deutschland brasilien asien usa europa indien frankreich china griechenland
Garbage	political conservative society economic religion republican cultural progressive movement politic politik gesellschaftlich partei politisch demokratie interessen wirtschaft ideologie politische demokratisch
Garbage	healthy healthful sustainable nonorganic conventional nongmo cheap nutritious fresh local regional umweltfreundlich hochwertige hochwertig nachhaltig regionale herkömmlich preiswert konventionell ökologisch
Garbage	often various typically certain generally numerous mostly largely frequently primarily häufig meist oftmals oft hauptsächlich manch zahlreiche zahlreich divers überwiegen
Garbage	etc via within toward towards throughout despite onto among across inklusive sowie aufgrund neben außerhalb samt bezüglich innerhalb etc wegen
Garbage	put turn pull bring throw blow move toss hang run schieben packen holen kippen ziehen werfen fallen setzen drücken legen
Garbage	festival restaurant dining cafe theater brewery studio music cocktail dance besucher restaurant restaurants gäste kulinarisch frühstück münchen laden küche urlaub
Garbage	difficult important necessary effective relevant helpful useful beneficial appropriate dangerous sinnvoll fragwürdig wichtig problematisch notwendig bedenklich schwierig wirksam unbedenklich nötigen
Garbage	certainly definitely therefore obviously simply generally still somewhat completely always trotzdem sicherlich durchaus dennoch prinzipiell oftmals grundsätzlich keineswegs absolut
Garbage	nearly almost although virtually unfortunately currently though still yet however ohnehin bislang bisher mittlerweile längst leider inzwischen fast derzeit allerdings
Garbage	unique diverse simple perfect different beautiful wonderful great popular comfortable simpel toll angenehm schöne besonder belieben wunderbar ideal komplex modern
Garbage	recognize acknowledge demonstrate tell understand conclude explain believe imply assume beweisen bemerken behaupten unterstellen bestreiten vorwerfen erwarten befürchten erwähnen feststellen
Garbage	something anything someone really anyone everyone nobody anybody everybody everything sowas wirklich jemand quatsch irgendwas blödsinn niemand irgendwie wozu sorry
Garbage	sustain improve extend achieve provide apply contribute maintain create ensure fördern unterstützen stoppen steigern berücksichtigen erzielen verhindern finanzieren festlegen anwenden
Garbage	tell ask talk listen understand complain inform worry learn hear beschweren trauen wünschen aufregen wundern jammern erwarten informieren erfahren denken

Table 13 Clusters from supervised method for GloVe Embeddings

Topics	Top 10 words for English and German
Biology	fish fruit animal wild feed eating insect usually tree bird tiere pflanzen fressen insekten ernähren arten nahrung tieren wachsen fische
Food & Cooking	sauce chicken cheese salad pasta soup bread butter dish tomato zutaten gemüse kochen kuchen speisen butter kartoffeln braten küche servieren
Food	fruit milk juice sugar corn bread meat coffee cheese vegetable gemüse kartoffeln zucker kaffee milch obst zutaten wein tomaten getreide
Farming & Livestock	livestock farming cattle animal agricultural dairy poultry crop breeding agriculture rinder schweine schafe nutztiere landwirtschaft gemüse getreide kühe geflügel schweinen
Farming	toxic pesticide bacteria harmful acid ingredient dietary chemical contamination protein substanzen bakterien inhaltsstoffe chemikalien stoffe lebensmitteln kohlenhydrate medikamente nährstoffe nahrungsmitteln
Work	job employee worker pay quit money hire earn leave staff verdienen job nebenbei angestellt jobs arbeitslos bezahlen geld mitarbeiter arbeit
Illness	disease illness cancer infection treat brain cause patient risk chronic krankheiten patienten krankheit behandlung ursache auftreten verursachen medikamente fällen krebs
Family Role	mother father life love wife husband couple friend woman man frau mutter vater kind eltern freund kinder tod junge allein
Shop	grocery vegetarian supermarket vegan dairy grower gourmet grocer healthful seafood vegane supermärkten veganer vegetarische discounter veganen supermärkte milchprodukte vegetarisch alnatura
Month	first last june since march july october september world april anfang september ebenfalls mai november oktober bereits juni märz februar
Color	look dark usually instead blue rather bright white black shape oben schwarz farbe ähnlich augen unten meist kopf manchmal weiß
Trend	increase amount cost income pay revenue less raise growth value kosten erhöhen investitionen dafür dadurch dollar millionen wert somit milliarden
Number	last week three five six next first month four eight anfang bereits ende nachdem zuvor zehn fünf drei sechs vier
Shop	walmart costco safeway retailer halo trump store supermarket omega chain regal bioware mist ehemalig hennen amazon marke label läden geschmacksverstärker
Business	company business sell retail industry marketing consumer manufacturer brand sale produkte unternehmen kunden firmen produkten verkauf geschäft firma umsatz anbieter
Job & Name	smith john james director michael robert david peter author paul schmidt müller koch fischer michael hans werner peter bauer autor
Education	education study research science development social institute environmental scientific focus forschung wissenschaft arbeit entwicklung soziale insbesondere bildung bereichen themen
Location	building nearby room downtown outside hotel restaurant apartment garden shop haus restaurant befinden wohnung straße heute ecke anlage neben hof
Location	area near along east part town north northern southern west norden westlich westen gebiet nahe richtung osten grenze etwa rund
Politics	democratic republican election democrats candidate democrat party vote senate conservative partei wahl mehrheit parteien politik politische parlament stimmen spd regierung
Politics	government administration must public commission security decision policy law federal behörden dafür deshalb maßnahmen weiterhin insbesondere jedoch gleichzeitig allerdings ebenso
Garbage	really pretty something maybe stuff everybody happy stupid fun crazy wirklich glücklich bisschen schön traurig bloß spaß wunderbar lustig warum
Garbage	water surface heat dry temperature fuel cold gas enough natural wasser luft temperatur strom dadurch erde entstehen relativ dabei deshalb
Garbage	starve hungry consume suck ingest feed tend educate clothe anymore füttern gebären verhungern fressen wärmen meiden aufwachsen küken brauchen ernähren
Garbage	back away put let keep turn instead pull hand able bringen ziehen nehmen fahren holen bewegen lassen werfen schlagen legen
Garbage	something really anything certainly else nothing sure maybe everyone actually wirklich gar niemand niemals tatsächlich jemand trotzdem weder genauso dennoch
Garbage	socalled andor angele etc bittman specie blvd lol csa ave verdenken biolebensmittel draufsteht biogemüse nunmal biobetriebe bioanbau jeglich koennen sysop
Garbage	intend might ask believe let able anyone agree ignore anything akzeptieren müssten lassen zulassen ignorieren ablehnen rechtfertigen erklären begründen unbedingt
Garbage	able enough allow reduce sell create generate help require cheap herstellen transportieren sparen verarbeiten produzieren erzeugen verwenden benötigen lassen reduzieren
Garbage	therefore certain rather particular necessarily appropriate must specific example instance andererseits grundsätzlich vielmehr wichtig möglich einerseits sinnvoll insofern deshalb berücksichtigen
Garbage	though however might yet able although fact rather likely must dennoch deshalb allerdings obwohl jedoch trotzdem sogar weshalb dadurch dafür
Garbage	fact indeed reason kind rather particular nothing believe understand sense vielmehr andererseits tatsache einerseits bezug meinung weder sinne aussagen gedanken
Garbage	available web internet information online example instance provide application addition möglich beispielsweise möglichkeit internet verfügung verfügbar informationen anwendung ermöglichen zusätzlich
Garbage	able help allow must continue bring encourage seek give provide lassen fördern unterstützen nutzen ermöglichen setzen möglichkeit helfen entwickeln
Garbage	although though example rather often however usually fact generally especially ebenso sowohl beispielsweise oft allerdings häufig meist gegensatz deshalb obwohl

Table 14 Clusters from unsupervised method for GloVe Embeddings

Topics	Top 10 words for English and German
Farming	crop corn wheat grain farming soybean agricultural harvest vegetable fruit kartoffeln mais gemüse getreide obst weizen saatgut anbau bananen angebaut
Animal & Livestock	animal dog cow pig meat cattle fish goat livestock eating tiere schweine rinder schafe tieren hunde fressen katzen kühe geflügel
Cooking	mixture stir butter sauce cooking boil oven combine cool spoon kochen trocken backen weich wasser kalt geschmack braten speisen waschen
Food	cheese sauce tomato bread butter dessert salad chicken cream juice zutaten gemüse kartoffeln butter milch käse geschmack tomaten zucker kaffee
Nutrition	dietary vitamin sodium nutritional acid cholesterol nutrient fatty protein diet inhaltsstoffe kohlenhydrate substanzen eiweiß vitamin vitamine nährstoffe pflanzliche lebensmitteln stoffe
Illness	disease infection illness cancer cause treat risk immune chronic medication krankheiten verursachen patienten ursache behandlung krankheit medikamente auftreten bakterien auslösen
Organ	skin stomach brain heart bone eye blood mouth pain muscle augen mund nase haut körper gesicht kopf gehirn blut auge
Production	fuel supply manufacture production waste equipment gas quality export amount produkte herstellung produkten hergestellt herstellen lebensmitteln produzieren erzeugung kohle rohstoffe
Family Role	mother father husband wife child daughter couple woman girl friend mutter frau vater eltern kind kinder freund kindern tod hause
Material	plastic glass thick metal thin piece wood leaf bag bottle glas schale durchmesser hergestellt meist früchte farbe manchmal herstellung kupfer
Resource	page web website news magazine online article newspaper blog interview artikel interview online zeitung internet bericht spiegel informationen medien veröffentlichen
Education	graduate university study science college professor student school education teacher universität studium medizin institut ausbildung forschung studenten mitarbeiter arbeit schüler
Shop	grocery store supermarket restaurant shop bakery convenience shopping retail chain supermarkt restaurants läden supermärkte filialen discounter restaurant geschäfte geschäft supermärkten
Traffic	bus train road along route traffic ride stop across car fahren richtung unterwegs vorbei täglich strecken verkehr kilometer gefahren minuten
Trend	increase amount cost income less pay value rate revenue rise kosten millionen dollar erhöhen milliarden umsatz prozent wert investitionen dadurch
Location	inside instead look back dark hand front turn rather side oben unten meist ebenso dabei außen kopf wobei augen ebenfalls
Job & Name	smith james john baker robert scott david michael jim mike müller schmidt klaus fischer koch werner wolfgang bauer hans ernst
Place	area nearby near water along land forest lake mountain river umgebung gegend liegen gebiet nahe wald westlich etwa norden fläche
Region	country europe western part region south united nation european southern europa ländern staaten frankreich deutschland sowohl anfang länder italien insbesondere
Business	company business sell industry marketing corporate investment firm offering purchase unternehmen kunden firmen produkte firma anbieter verkauf konzern hersteller produkten
Politics	democratic republican democrats election senate candidate democrat vote party clinton wahl partei parteien spd mehrheit politik parlament cdu regierung bundestag
Education & Politics	education public government development environmental health organization provide program protection organisation maßnahmen zusammenarbeit insbesondere initiative forschung arbeit weiterhin unterstützung bildung
Garbage	something really anything certainly nothing else actually indeed sure always gar wirklich dennoch tatsächlich trotzdem obwohl deshalb weder niemand offensichtlich
Garbage	socalled nonorganic foodie contaminate organically healthful longterm angele nutritious vegan verdenken nunmal jeglich biogemüse biolebensmittel biobranche biobetriebe draufsteht bioanbau bioprodukten
Garbage	therefore example particular instance certain specific rather thus must appropriate möglich beispielsweise grundsätzlich deshalb möglichkeit andererseits entweder verwenden entsprechend somit
Garbage	last though although however three first next since later month anfang bereits nachdem ebenfalls ende zuvor jedoch schließlich allerdings dabei
Garbage	example music original contemporary famous though kind part modern history ebenso ebenfalls sowohl neben bereits anfang bekannt beispielsweise außerdem dabei
Garbage	fact whether reason indeed yet question might believe nothing possible tatsache tatsächlich allerdings aussagen zusammenhang deshalb meinung offensichtlich jedoch dennoch
Garbage	let understand whatever able intend anything ignore might tell surely begreifen müssten bemerken akzeptieren lassen ignorieren vorstellen anfangen hoffen zweitens
Garbage	first second next win final third last play start place jeweils gewinnen spielen dabei saison erneut zweite nachdem allerdings zudem
Garbage	really pretty maybe something stupid fun crazy stuff guess everybody glücklich wirklich schön bisschen traurig genug bloß wunderbar irgendwie spaß
Garbage	able might must bring keep let help try enough allow bringen verhindern deshalb nehmen lassen trotzdem versuchen dennoch setzen sofort
Garbage	rather particular kind sense fact indeed understanding context focus understand andererseits einerseits vielmehr sinne bezug verständnis insbesondere sicht ideen dennoch
Garbage	allow able must encourage help seek continue intend decide give lassen unterstützen fördern setzen möglichkeit kontrollieren müssten durchführen andererseits organisieren
Garbage	though rather although often generally especially usually fact less quite ebenso allerdings oft deshalb kaum sowohl obwohl hingegen dennoch meist