Word Embedding-based Task Adaptation from English to Hungarian

Zsolt Szántó, Carlos Ricardo Collazos García, Richárd Farkas

University of Szeged, Institute of Informatics
Árpád tér 2., Szeged, Hungary

Black Swan Data Inc.
Tisza L. krt 47., Szeged, Hungary

{zsolt.szanto, richard.farkas}@blackswan.com

Abstract: In commercial Natural Language Processing (NLP) solutions, we frequently face the problem, that a particular NLP application has to work on several languages. Usually the solution is first developed on a single language – the *source* language – then it is adapted to the other languages – the target languages . In this paper, we introduce experimental results on English to Hungarian adaptation of document classification tasks. In our setting, only an English training dataset is available and our aim is to get a classifier which works on Hungarian documents. We experimented comparatively with two different approaches for word embedding-based language adaptation methods and evaluated them along with monolingual methods in a sentiment classification and a topic classification dataset.

1 Introduction

In commercial Natural Language Processing (NLP) solutions, we frequently face the problem, that a particular NLP application has to work on several languages. Usually the solution is first developed on a single language – the *source* language hereafter – then it is adapted to the other languages – the target languages. There are many opportunities for these adaptations:

1. The necessary resources (like training data labeling and dictionaries) can be constructed manually from scratch to the target language following the principles and best practices recognized during the experiments on the source language. Then we can train models exploiting the brand new resources.
2. The necessary resources can be translated from the source language to the target language then we can train models on these translated resources. Translation can be done manually or by machine translation systems. In both cases it might introduce errors like for example the translation of dictionary items without knowing their purpose/context might be problematic for humans as well.

3. Statistical approaches can be applied for language adaptation itself. Adaptation in this case is not text translation but for example it can be carried out in word embedding spaces.

In this paper, we introduce experimental results with the 3rd approach on English to Hungarian adaptation of document classification tasks. In this setting, only an English training dataset is available and our aim is to get a classifier which solves the same task on Hungarian documents. We comparatively experimented with two different approaches for word embedding-based language adaptation methods and evaluated them, along with monolingual methods, in a sentiment classification and a topic classification dataset.

To the best of our knowledge, this is the first work on automatic language adaptation of any NLP tasks to Hungarian.

2 Word Embedding-based Language Adaptation Techniques

There have been two main approaches published for word embedding-based language adaptation. The earlier approach utilizes a bilingual dictionary to train a mapping between the monolingual word embeddings of the source and target languages [6]. The recent approaches exploit parallel corpora and construct a bilingual word embedding from it [11]. Here, we briefly introduce the principles of word embedding and these two approaches for language adaptation.

2.1 Word Embedding

A word embedding is a distributional representation of words in a few hundred dimensional continuous vector space [7], [10]. Two vectors are close to each other in the embedding space if the words they belong to are similar to each other. More precisely, if two words appear in similar contexts their vector representations are pushed to be close to each other during the construction of the word embedding.

The distributed representations for words have become extremely successful. Their main advantage is that they can help to model unseen or rare words. Usually we train a word embedding on huge unlabeled corpora. On the other hand, the training corpus in a supervised machine learning setting is relatively small. In prediction time, if we find a word which was not present in the training data we can look for similar words in the word embedding thus generalizing the patterns learnt from the training data.

Mikolov et al. [8] also showed that the distributed representations of words capture surprisingly many linguistic regularities, and that there are many types of similarities among words, which can be expressed as linear translations.

There are two popular models for learning word embedding efficiently on large amounts of texts, namely Skip-gram and CBOW [5]. Here, we briefly introduce Skip-gram as it is extended into a bilingual model we use in this paper. In the Skip-gram model [5], the training objective is to learn word vector representations that are good at predicting their context in the same sentence. More formally, given a sequence of

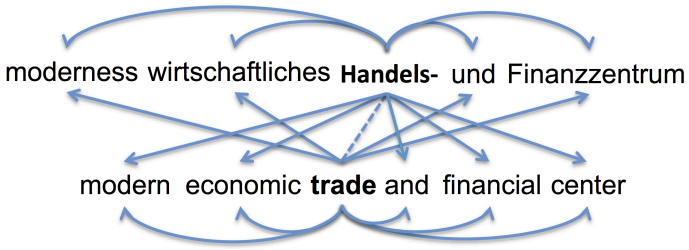

Figure 1. A German and English word aligned phrase to depict the BiSkip model. It exploits monolingual context like skip-gram and also cross-linguality based on the given word alignment [4].

training words, the objective of the Skip-gram model is to maximize the average conditional
probability of the words in a given window conditioned on the middle word. In the Skip-gram model, every word is associated with two learnable parameter vectors, the word vector and the context vector. After training, the context vectors are dropped and word vectors are used as word embedding.

Due to its low computational complexity, the Skip-gram model can be trained on a large corpus in a short time, i.e. billions of words in hours.

2.2 Bilingual Dictionary-based Adaptation

The first attempts at word embedding-based language adaptation were based on two pretrained monolingual word embeddings on both the source and the target languages. Then they learn a mapping, called translation matrix, between the two vector representations exploiting the entries of a bilingual dictionary as training examples [6].

The hypothesis behind this approach is that the vector representations of similar words in different languages are related by a linear transformation. Hence they optimize for a *translation matrix,* which is able to linearly transform any words from the target language's word vector space to the source language's vector space with minimum distance.

For a document classification task, we can train a machine learning model on the available English training corpus utilizing the word vectors from the document. In prediction time, we look for each word vector of the Hungarian document's words, map each of them through the translation matrix into the English word embedding space. Then the machine learnt model makes its prediction on the translated vectors.

2.3 Parallel Corpus-based Adaptation

More recent approaches aim to learn a single joint bilingual word embedding for the source and target languages from a parallel corpus [11]. Their assumption is that by allowing the joint model to utilize both the co-occurrence context information within a

language and the meaning-equivalent signals across languages, they can obtain better word vectors both monolingually and bilingually. We use the so called *BiSkip* bilingual embedding model in this paper because in Upadhyay et al. [11], BiSkip proved to be the most robust and accurate in comparison with other state-of-the art bilingual embedding models.

Luong et al. [4] proposed the Bilingual Skip-Gram (BiSkip) algorithm, an extension of the monolingual skip-gram model, which learns bilingual embeddings by using a parallel corpus along with word alignments (see Figure 1). The learning objective is an extension of the skip-gram model, where the context of a word is expanded to include bilingual links obtained from word alignments, so that the model be trained to predict words cross-lingually.

Figure 2 shows colors in English – Hungarian bilingual vector space. We used PCA to reeducate the dimensions of the vectors.

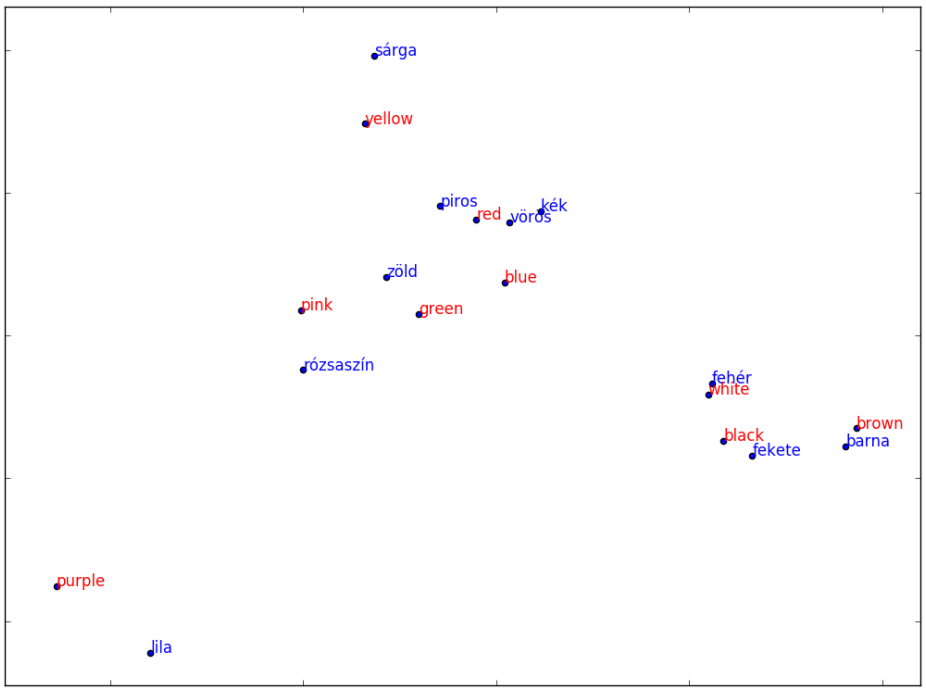

Figure 2. Hungarian and English colors in BiSkip trained vectors.

3 Evaluation Datasets

We evaluated the adaptation approaches in two types of classification tasks, sentiment and topic classification. In each case, we worked on user generated short texts from social media. Both datasets are binary classification problems (i.e. there are two class

labels) and the distribution of the labels is uniform. The sentiment corpora have *positive* and *negative* labels, while the topic classification task contains *game* and *sport* labels.

Table 1 summarizes the sizes of the train and the evaluation datasets for English and Hungarian for the two evaluation scenarios.

Table 1. Sizes of datasets used for evaluating language adaptations.

	Sentiment		topic	
	HU	EN	HU	EN
train #doc	5 000	5 000	10 000	10 000
eval #doc	1 000	1 000	2 000	2 000
train #token	60 945	10 5648	173 655	153 662
eval #token	12 122	20 307	33 274	29 338

3.1 Sentiment Classification

We downloaded product reviews from the English `newegg.com` and the Hungarian `arukereso.hu` sites. The reviews are coming from the IT domain in both languages. Moreover, both sites contain *pro* and *con* fields where a user summarizes his opinion. We used only these summaries and took *pro* as *positive* and *con* as *negative* documents. We removed the too short (less than 4 tokens) documents as they usually hold only placeholder content, like 'none'.

3.2 Topic Classification

In topic classification, our objective was to develop a classifier, which is able to identify the central topic of any user-generated text (like Facebook posts or tweets). For a feasibility study, we downloaded public `Facebook` posts from Hungarian and English sites in the *computer/console game* and *sports* topics through the Facebook Graph API[1]. These Facebook sources are listed in Table 2.

Table 2. The sources of topic classification datasets (Facebook pages)

HU	game	PCGuruMagazin, gamestarhu, 576Kbyte, gamedayiroda
HU	sports	nsonline, focihiradohu
EN	game	kontaku, pcgamemagazin
EN	sports	SkySports, ESPN

[1] https://developers.facebook.com/docs/graph-api/

4 Experimental Setting

In this section, we describe the bytes and bits of our experimental setups for the two adaptation approaches.

4.1 Translation matrix

For the translation matrix approach, we employed the Polygot pretrained embeddings [1]. Polyglot embeddings are publicly available for more than 100 languages trained on Wikipedia dumps of the languages (89 million Hungarian and 1704 million English tokens) using the Skip-Gramm model. Each polyglot model contains the most frequent 100000 words from the selected dump. We used their 64 dimensional vectors for Hungarian and English. The mapping between the embeddings of the two languages was learnt on the Universal dictionary database[2], i.e. we were optimizing a mapping which can achieve the minimum of the sum squared error on mapping Hungarian word vectors of the dictionary to English word vectors. We trained a linear regressor for each of the dimensions and also experimented with Canonical-correlation Aanalysis but they could achieve similar results.

Having the word vector mapping, we train a classifier on the English training dataset then in prediction time, we map the word vectors of the Hungarian document in question into the English word embedding space and carry out the classification based on the mapped vectors. In our experiments, we calculated the average of the word vectors of a document and use these averages as features of a logistic regression classifier (using the `python sklearn` implementation with its default metaparameters [9]). We also tried neural network based approaches, Convolutional Neural Network and Recurrent Neural Networks for exploiting the word vector representations but could not get higher scores.

4.2 BiSkip adaptation

The bilingual word vectors were constructed on 10 million English-Hungarian sentence pairs of the `OpenSubtitles` parallel corpus [3]. We chose this parallel corpus for our experiments because movie subtitles are closer to social media texts than other available parallel corpora as they use more slang and have a conversational nature. First, we calculated word alignment on the parallel corpus with `fast_align`[3] [2] then we trained the BiSkip bilingual word embedding with the `bivec`[4] tool [4], (we used the default parameters, dimension: 200, window size: 5, iterations: 50). The bilingual model contains 298728 Hungarian and 120615 English words.

In this scenario, we train again a logistic regression classifier on the English training dataset using the average of the word vectors as document features. In prediction time,

[2] http://www.dicts.info/uddl.php
[3] https://github.com/clab/fast_align
[4] https://github.com/lmthang/bivec

we take the average of the Hungarian word vectors and apply the classifier on the top of them as they are consistent with the English embedding.

Table 3 shows the ratio of words from the classification datasets which are not in the vocabularies of the word embeddings.

Table 3. Ratio of the out-of-vocabulary words in classifications datasets.

	Sentiment		topic	
	HU	EN	HU	EN
polyglot	10.26%	6.32%	17.70%	8.07%
subtitles	8.00%	5.12%	15.53%	6.76%

5 Results

The baseline in each evaluation setting is 50% accuracy as we always have uniform label distribution and binary classification tasks. We compare the results of word embedding-based adaptation against monolingual results, i.e. against the accuracies that might be achieved if a Hungarian training corpus with the same size would be available. These results can be considered as upper bounds for the language adaptation scenario.

Table 4. Accuracies achieved by various models on two evaluation settings.

	sentiment	topic
bag-of-word	92.3	87.5
translation matrix	52.1	53.0
BiSkip, monoling, 10M sent	88.3	81.8
BiSkip, train on EN, 10M sent	80.1	66.2
BiSkip, monoling, 1M sent	87.1	76.6
BiSkip, train on EN, 1M sent	78.4	54.9

Table 4 consists of the accuracy scores achieved on the Hungarian evaluation datasets. The ‘bag-of-word’ and the ‘monoling’ models are trained on the Hungarian training dataset, hence they are upper bounds for the ‘translation matrix’ and ‘train on EN’ models which have access only to English training data. The bag-of-word model is a logistic regression classifier with uni- and bigram features. The word embedding-based approaches are introduced in the previous section.

We tried the following parameters: word vector sizes 50, 100, 200; number of iterations: 5, 20, 50. The table contains the best results among these parameter settings for each evaluation scenario.

6 Discussion

The monolingual results are an upper bound for the language adaptation experiments but there is a considerable gap between the two monolingual settings, i.e. between monolingual BiSkip and the bag-of-word results achieved. Both approaches train a logistic regression classifier on the Hungarian training dataset. The key difference between them is at the feature representation, which consists of uni- and bigram tokens versus average of word vectors. The reason for this gap may be the size of the training datasets as it might happen that few thousand training examples are sufficient to learn the contribution of particular uni- and bigrams and the average of word vectors becomes to be too general.

Table 3 shows that the translation matrix approach failed in these experimental setups. Most likely, the Polyglot word embedding – trained on the Wikipedia – is not suitable for the distributed representation of social media text. Another explanation might be that the one-to-one translation of words from Hungarian to English is not linear. For instance, it should map each form of a Hungarian noun to the same vector in the English embedding.

BiSkip could achieve much better results. Its bilingual results are fair to compare with its monolingual results as it avoids the bag-of-words versus vector representation effect. The difference between the monolingual and bilingual results are 8 and 15 percentage points on the sentiment and topic classification tasks, respectively. This is the price we have to pay if we do not label a monolingual training dataset but employ a state-of-the-art automatic language adaption technique. A possible reason for the difference between the gaps at the two tasks is that in topic classification named entities are more important than in the sentiment task, and the translation of the named entities are very easy, but if the parallel corpus (which created from subtitles) does not contain a named entity you cannot generate a word vector to it.

Finally, Table 3 also reveals the effect of data amount which the word embedding was calculated on. In each setting, a considerable improvement can be observed if BiSkip can be trained on 10 million sentence pairs instead of 1 million sentence pairs.

7 Conclusions

We introduced experiments with state-of-the-art language adaptation techniques from English to Hungarian. We assume a document classification task where only an English labeled training dataset is available but we aim to solve the same classification task in Hungarian documents. Our experiments on a sentiment and on a topic classification task showed that the translation matrix-based method failed while the BiSkip method could considerably outperform it. Our experiments support that the corpus which the word embedding is trained on and the document classification corpus have to be as lose as possible in domain and that the size of the parallel corpus exploited is important. Our final conclusion is that state-of-the-art language adaptation methods can achieve roughly 10 percentage point worse results compared to the situation where a labeled training corpus would be available.

Acknowledgements

We are grateful for the work of Viktor Hangya on downloading and cleaning the sentiment classification datasets.

The research of Richárd Farkas is supported by the János Bolyai Research Scholarship of the Hungarian Academy of Science.

References

1. Rami Al-Rfou, Bryan Perozzi, Steven Skiena: Polyglot: Distributed Word Representations for Multilingual NLP. In Proc. CoNLL, pp 198-192 (2013)
2. Chris Dyer, Victor Chahuneau, and Noah A. Smith: A Simple, Fast, and Effective Reparameterization of IBM Model 2. In Proc. of NAACL (2013)
3. Pierre Lison, Jörg Tiedemann: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proc. of LREC (2016)
4. Minh-Thang Luong, Hieu Pham, Christopher D. Manning: Bilingual Word Representations with Monolingual Quality in Mind. In Proc. of NAACL-HLT. pp 151-159. (2015)
5. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. (2013)
6. Tomas Mikolov, Quoc V. Le, and Ilya Sutskever: Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* (2013)
7. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean: Distributed representations of words and phrases and their compositionality. In Proc. NIPS 26, pp 3111–3119. (2013)
8. Tomas Mikolov, Scott Wen-tau Yih, and Geoffrey Zweig: Linguistic regularities in continuous space word representations. In Proc. of NAACL HLT (2013)
9. Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 12, pp. 2825-2830 (2011)
10. David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams: Learning representations by back-propagating errors. *Nature*, 323 (6088):533–536. (1986)
11. Shyam Upadhyay, Manaal Faruqui, Chris Dyer, Dan Roth: Cross-lingual Models of Word Embeddings: An Empirical Comparison. In Proc. of ACL. pp 1661–1670 (2016)