Artificial Intelligence 274 (2019) 133–150

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Toward any-language zero-shot topic classification of textual documents

Yangqiu Song [a,*], Shyam Upadhyay [b], Haoruo Peng [c], Stephen Mayhew [b], Dan Roth [b]

[a] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
[b] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[c] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 11 November 2017
Received in revised form 19 January 2019
Accepted 6 February 2019
Available online 13 February 2019

Keywords:
Multilingual text classification
Cross-lingual text classification
Zero-shot text classification
Semantic Supervision

ABSTRACT

In this paper, we present a zero-shot classification approach to document classification in any language into topics which can be described by English keywords. This is done by embedding both labels and documents into a shared semantic space that allows one to compute meaningful semantic similarity between a document and a potential label. The embedding space can be created by either mapping into a Wikipedia-based semantic representation or learning cross-lingual embeddings. But if the Wikipedia in the target language is small or there is not enough training corpus to train a good embedding space for low-resource languages, then performance can suffer. Thus, for low-resource languages, we further use a word-level dictionary to convert documents into a high-resource language, and then perform classification based on the high-resource language. This approach can be applied to thousands of languages, which can be contrasted with machine translation, which is a supervision-heavy approach feasible for about 100 languages. We also develop a ranking algorithm that makes use of language similarity metrics to automatically select a good pivot or bridging high-resource language, and show that this significantly improves classification of low-resource language documents, performing comparably to the best bridge possible.

1. Introduction

With the increasing growth of textual information on the Web, there is a great need to determine the topics of each document. Many applications, including news classification [1,2], search result organization [3], online advertising [4], etc., have placed text categorization as a key problem. Compared to English text classification, classifying languages other than English can be challenging, since there is relatively less annotation for other languages. There have been several ways to handle text classification other than English. For example, cross-lingual document classification was proposed as a way to use training data in one language to classify the documents in another language [5,6]. This is useful when annotation is available in the source language and the cost of acquiring labeled data in the target language is high. However, when the topic space is changed, new annotation would be required for either source or target language to retrain the classifier.

* Corresponding author.
E-mail addresses: yqsong@cse.ust.hk (Y. Song), shyamupa@seas.upenn.edu (S. Upadhyay), hpeng7@illinois.edu (H. Peng), mayhew@seas.upenn.edu (S. Mayhew), danroth@seas.upenn.edu (D. Roth).

https://doi.org/10.1016/j.artint.2019.02.002
0004-3702/

Current research has not paid enough attention to the fact that the labels or short descriptions of the categories can be viewed as meaningful units of texts themselves. Given a collection of text documents and a set of categories, we show that it is possible to assign category labels to documents without requiring any labeled training data, by *understanding* the labels to accurately perform this categorization. This allows us to develop an on-the-fly classification procedure which requires only short descriptions of the categories, eliminating the need of labeled documents as supervision, making it essentially a zero-shot classification. Our approach is different from existing zero-shot learning [7–10] and one-shot learning [11,12] approaches, in that one-shot learning requires one example for training, while in zero-shot learning, the test data is different from the training data (e.g., a new label space). We call our approach cross-lingual zero-shot classification (CLZSC).

CLZSC embeds both labels in English and documents in another language into a semantic space that allows one to compute meaningful semantic similarity between a document and a potential label. We used two types of embeddings to achieve this goal. First, we use cross-lingual explicit semantic analysis (CLESA) [13,14] to map both English labels and documents in another language into the same space represented by Wikipedia concepts. CLESA aligns Wikipedia pages with the same title across languages using language links. By working in this aligned space, CLESA embeds texts in two languages into the same semantic space. In the second approach, we use cross-lingual word embeddings [15,16] as an alternative to CLESA. By averaging all words appearing in the labels and the documents using cross-lingual word embedding, we can simply compare the similarity based on the embedding vectors. We compared two approaches and show that CLESA is better when both the documents are clean and the label descriptions are adequate, while cross-lingual embeddings are better when the documents are noisy and the label descriptions are short. Both representations suffer when there are not enough resources to compute the shared semantic space across languages. This limits the application of such approaches to more languages in the world.

We demonstrate this low-resource problem using CLESA, since in our experiments, CLESA outperforms cross-lingual embedding on the multilingual text classification dataset we have for 88 languages. CLESA relies on mapping the English labels or short category description into a Wikipedia-based semantic representation, and on the use of the target language Wikipedia. Consequently, performance could suffer when Wikipedia in the target language is small. We tackle the challenge for languages with little or no presence in Wikipedia, which we refer to by *small-Wikipedia languages* (**SWLs**). One can think of multiple ways to facilitate classifying documents in SWLs into English categories. The simplest way is to translate the SWL documents to English or a language with large Wikipedia presence (*large-Wikipedia language*, **LWL**), and then apply English zero-shot classification or cross-lingual zero-shot classification for the LWL. Unfortunately, this requires full document translation which, in turn, requires large amounts of parallel data in the two languages to train a machine translation system. This is unlikely to be available anytime soon. The available resources for training standalone machine translation tools are relatively very sparse. For example, Europarl[1] covers 21 languages; Google Translate[2] covers 103 languages.

We first show that bi-lingual dictionaries (or "word-level translations") can be used to support reliable document classification via zero-shot classification. This approach scales to many languages due to existence of resources like PanDictionary [17] or later Panlex[3], which have word mappings for *thousands of languages*. Second, we show how to choose the best LWL to serve as the bridge language between a given SWL and English. For example, for Hausa, a SWL, it turns out that if we can find a related LWL, such as Arabic, then we can use the Arabic–English Wikipedia to perform CLZSC. Since Arabic is more similar to Hausa compared to English to Hausa, mapping of words from Hausa to Arabic can be better than English. While the idea of using a bridge (pivot) language is not new [18], in this paper we systematically evaluate CLZSC using 88 languages, including 39 SWLs and 49 LWLs, and show that this bridging approach successfully supports good classification of a large proportion of SWLs we tested. We also propose an automatic way to rank LWLs based on their ability to support good categorization of SWL documents. Specifically, we show how to use RankSVM [19,20] to learn from the language features to identify which LWLs should be effective as a bridge to a given SWL. Experiments show that this learning based method is significantly better than the use of hand-crafted language similarities to rank the LWLs, and that, in many cases, it selects the best possible bridge.

The contributions of this paper can be highlighted as follows:

- We propose a cross-lingual zero-shot classification framework which does not require labeled data. The classification can classify a document in another language into categories that can be represented as English label names or descriptions.
- We show strong performance of our classification framework for both English and many other languages using benchmark datasets. Our framework can achieve classification performance equal to supervised classification with hundreds of annotated documents per category.
- For the languages with small Wikipedia presence, we also proposed a bridged cross-lingual classification framework, using a third language which is related to both English and target language, to enrich the resources to have a better representation. We use a created multilingual dataset based on 20-newsgroups and Google Translate to test the performance. We also show that by using a multilingual dictionary such as Panlex, we can generalize our framework to thousands of languages.

[1] http://www.statmt.org/europarl/.
[2] https://translate.google.com.
[3] https://panlex.org/.

This paper is an extension of the previously published IJCAI 2016 paper [21] and an unpublished Arxiv manuscript [22]. Code for this paper is available at http://cogcomp.org/page/software_view/DatalessHC.

2. Zero-shot classification framework

We first introduce the general framework of the zero-shot classification framework and then show two different representations of labels and documents. Our classification scheme consists of two steps: the first is an initial on-the-fly zero-shot classification and the second performs bootstrapping.

2.1. On-the-fly zero-shot classification

We perform a nearest neighbor search of labels for a document in an appropriately selected semantic space [23,24]. Let $\phi(d)$ be the representation of document d in a semantic space (to be defined later) and let $\{\phi(l^{(1)}), \ldots, \phi(l^{(L)})\}$ be the representations of the L labels in the same space. Then we can evaluate the similarity using an appropriate metric $f(\phi(d), \phi(l^{(i)}))$, (e.g., cosine similarity between two sparse vectors) and select label(s) that maximizes the similarity:

$$l^* = \arg\max_i f(\phi(d), \phi(l^{(i)})). \tag{1}$$

The core problem in zero-shot classification is to find a semantic space that enables good representations of documents and labels. Traditional text classification makes use of a bag-of-words (BOW) representation of documents. However, when comparing labels and documents in zero-shot classification, the brevity of labels makes this simple-minded representation and the resulting similarity measure unreliable. For example, a document talking about "sports" does not necessarily contain the word "sports." Consequently, other more expressive distributional representations have been applied, e.g., Brown cluster [25,26], neural network embedding [27–30], topic modeling [31], ESA [32], and their combinations [33]. Among different representations, it has been shown that ESA gives the best and most robust results for zero-shot classification for English documents [24]. ESA uses Wikipedia as external world knowledge to generate a set of *titles* for a given fragment of text [32]. Each word in a text is represented as a weighted vector of the Wikipedia titles in which it is mentioned. This can be computed using an inverted index for each word in Wikipedia. The text fragment representation is then the sum of the IDF (inverse document frequency) weighted vectors that correspond to the words in the text fragment.

2.2. Bootstrapping

We also use a bootstrapping procedure for the zero-shot classification. This is a natural step to follow since it is free (no labeled data is needed) and it provides generic semantic representations to best fit the specific data collection. The bootstrapping step makes use of unlabeled data (the given document collection or additional unlabeled data if so desired), and it labels the most confident documents in each iteration, starting with the labels given in the on-the-fly zero-shot classification. Then, it trains a new classifier to improve its accuracy and incorporate more labeled data. The procedure is as follows:

Step 1: Initialize N_0 documents for each label, using confident on-the-fly zero-shot classifications.
Step 2: For each iteration, train a classifier based on BOW representation to label N more documents for each label.[4]
Step 3: Continue until no unlabeled documents remain.

2.3. Cross-lingual document representation

Choosing the right semantic representation $\phi(x)$ is crucial to performance for any downstream task. In this section, we propose two possible representations.

2.3.1. Cross-lingual ESA (CLESA)

We first show how we can easily extend our framework to handle cross-lingual classification tasks by representing concepts in different languages in the same semantic space. In order to support cross-lingual zero-shot classification, we implemented a version of CLESA [13,14] that is used for zero-shot classification scheme by exploiting the shared semantic space between two languages. CLESA is a generalization of explicit semantic analysis (ESA) for English [32], introduced in the context of Information Retrieval [13,14] and used also for Twitter message classification [34]. To build connections between languages, we extract cross-language links from Wikipedia dumps of X languages. Each such link identifies a pair of corresponding titles in two different languages. For example, a Wikipedia page titled "Basketball" has a corresponding Italian page "Pallacanestro," a Spanish page "Baloncesto," etc. Note that though the titles need not be direct translations, they define the same semantic concept.

[4] Our experiments show that bootstrapping with BOW features is the best choice among the different semantic representations.

Using these cross-language links, we can intersect the Wikipedia title space of any two languages and use the set of shared Wikipedia titles as the *shared semantic space* for texts in both languages. Formally, assume that we have Wikipedia dumps for languages A and B. Traditional ESA uses the sparse vector $\phi^A(w_A) = (\phi^A_{C_{1_A}}(w_A), \ldots, \phi^A_{C_{N_A}}(w_A))^T \in \mathbb{R}^{N_A}$ to represent a word w_A where N_A is the number of titles in the language A Wikipedia, and $\phi^A_{C_{i_A}}(w_A)$ is the weight indicating how important word w_A is in the Wikipedia page titled C_i in language A. Similarly, $\phi^B(w_B) = (\phi^B_{C_{1_A}}(w_B), \ldots, \phi^B_{C_{N_B}}(w_B))^T \in \mathbb{R}^{N_B}$ for language B. To compare text similarities between languages A and B, a natural way is to consider first the intersection of the two title sets:

$$\{C_1, \ldots, C_N\} = \{C_{1_A}, \ldots, C_{N_A}\} \bigcap \{C_{1_B}, \ldots, C_{N_B}\}. \tag{2}$$

Thus, we unify the vector representations in A and B as follows:

$$\begin{aligned} \phi(w_A) &\doteq (\phi^A_{C_1}(w_A), \ldots, \phi^A_{C_N}(w_A))^T \in \mathbb{R}^N, \\ \phi(w_B) &\doteq (\phi^B_{C_1}(w_B), \ldots, \phi^B_{C_N}(w_B))^T \in \mathbb{R}^N. \end{aligned} \tag{3}$$

Now, suppose we have a document d_A in language A as a vector $(w_{A_1}, \ldots, w_{A_{M_A}})^T \in \mathbb{R}^{M_A}$, where M_A is the vocabulary size of language A. Denote p_{A_i} as the weight of word w_{A_i} in the document. For example, the weight could be TF-IDF, where TF represents the term frequency of word w_{A_i} in d_A, and IDF, the inverse document frequency in Wikipedia. Then we can define the vector representation for d_A as:

$$\phi(d_A) = \frac{1}{M_A} \sum_i p_{A_i} \phi(w_{A_i}). \tag{4}$$

Similarly, for a label $l^{(i)}_B$ in language B, we have the vector representation:

$$\phi(l^{(i)}_B) = \frac{1}{M_B} \sum_j p^{(i)}_{B_j} \phi(w^{(i)}_{B_j}), \tag{5}$$

where $l^{(i)}_B = (w^{(i)}_{B_1}, \ldots, w^{(i)}_{B_{M_B}})^T \in \mathbb{R}^{M_B}$ is a highly sparse vector, M_B is the vocabulary size of language B and $p^{(i)}_{B_j}$ represents the weight of word j in the label description $l^{(i)}_B$. Now we can use the cosine similarity between $\phi(d_A)$ and $\phi(l^{(i)}_B)$ as in traditional ESA in order to choose the best label:

$$l^* = \arg\max_i \cos(\phi(d_A), \phi(l^{(i)}_B)). \tag{6}$$

2.3.2. Cross-lingual embeddings

We also use cross-lingual word embedding to support cross-lingual zero-shot classification. Motivated by the simplicity and success of neural network-based word embedding [29,30], multilingual [35] or cross-lingual [6,36–40] representation learning are also investigated. Similar to cross-lingual classification, traditional representation learning approaches need either parallel corpora [6,37], some labeled data in the target domain [36], or words being (partially) aligned in a dictionary [41]. More recent studies have also shown that we can first train two sets of monolingual embeddings in different languages and then align them either using a seed parallel lexicon [15] or without employing any cross-lingual annotated data [16]. Given that cross-lingual word embeddings can be in a same embedding space, we also follow previous approach [24] by using a weighted average of all word embeddings in a document to represent the document as a vector in the same space. Specifically, we follow the scheme used for ESA to generate the document semantic representation using a TF-IDF weighted combination of word vectors in the documents. Formally, we use the following representation for a document:

$$\psi(d_A) = \frac{1}{M_A} \sum_i p_{A_i} \psi(w_{A_i}) \tag{7}$$

and a label:

$$\psi(l^{(i)}_B) = \frac{1}{M_B} \sum_j p^{(i)}_{B_j} \psi(w^{(i)}_{B_j}), \tag{8}$$

where the definitions for M_A, M_B, p_{A_i}, p_{B_i}, w_{A_i}, $w^{(i)}_{B_j}$, d_A, and $l^{(i)}_B$ are all the same as CLESA for languages A and B. The only difference is that $\psi(\cdot)$ is a dense vector representation as word embedding.

2.4. Bridging for small Wikipedias

There are more than 7,000 known spoken languages, and 3,000 of them have writing systems.[5] Among them, only hundreds of them have Wikipedia. On the other hand, we have dictionaries for many more languages. Thus, we can use the dictionaries or lexicons to further extend the use of cross-lingual zero-shot classification. A natural question is that, given an SWL document, which bi-lingual dictionary should we use to facilitate good classification into English categories? Mapping to English may not be the best. For example, among the 169 languages in Wiktionary[6] we can download, there are more than 800 language pairs with more than 1,000 language links, but only 59 of them are associated with English. This analysis indicates that in order to facilitate classifying an SWL document into an English ontology, we may need to go through a *bridge* language, an LWL for which bi-lingual dictionaries are available.

When using different languages as a bridge language, we find some related languages in Latin writing system, such as Spanish, Catalan, Indonesian, and Bulgarian, are ranked high for bridging Hausa and Uzbek. Some of the top bridging LWLs are in the same family with the target languages. For example, Arabic, Hebrew, and Hausa are in the Afro–Asiatic family. Moreover, region of the native speakers is also reflected. Persian speakers in Iran live relatively near to Uzbek speakers in Uzbekistan. However, writing system, language family, and region are not the only factors that affects the ranking. Besides other linguistic factors, we also presume that either the size of Wikipedia or the less ambiguity of translation may help them result in relatively good accuracy. Then the remaining question is how to automatically select a good bridging LWL for the SWL document classification.

2.4.1. Ranking based on heuristics

To automatically rank the bridging LWLs for SWLs, we first use the World Atlas of Language Structures (WALS)[7] data as language features. At time of download, there were 2,679 languages with 198 features including phonological, grammatical, and lexical properties. We removed latitude and longitude features thus resulting in 196 features. Given the above analysis, we found four features that are very useful compared to the others, which are *genus*, *family*, *macro area*, and *country code*. From these, we manually developed a similarity value for a pair of languages as follows:

$$\begin{aligned} S_l(L_1, L_2) &= 50 \cdot I_{genus}(L_1, L_2) \\ &+ 50 \cdot I_{family}(L_1, L_2) \\ &+ 50 \cdot I_{macro\ area}(L_1, L_2) \\ &+ 50 \cdot I_{country\ code}(L_1, L_2) \\ &+ \textstyle\sum_{i \in \{others\}} I_i(L_1, L_2), \end{aligned} \tag{9}$$

where $I_i(L_1, L_2) = 1$ means that two languages L_1 and L_2 share the same feature i. If one of the most important features is identified, we add a large value into the similarity value. We set the weight to be 50 with the heuristics that the number can be comparable to the number of 196 features.

We also want to incorporate the size of the Wikipedias since it correlated well with the performance of CLZSC. Therefore, we rank the languages with the size of Wikipedias:

$$S_w(L) = \#\text{Wikipedia Title in } L. \tag{10}$$

If Wikipedia size is the only factor, we will always use English as the bridging language. Besides the Wikipedia size, we also use the language links to rank the bridging languages:

$$S_{ll}(L_1, L_2) = \#\text{Language links from } L_1 \text{ to } L_2. \tag{11}$$

To combine the two ranking factors, since the scales of two similarity/size values are different, we first convert each similarity/size value to the rank value. A larger score denotes a more highly ranked language. We use W_l and W_w as the weights for each similarity. For example, German is ranked as second by S_w, and there are 49 candidate languages, then $W_w(\text{German}) = (49 - 1)/49 = 0.980$. Then we use the harmonic mean of two weights as the combined rank value:

$$S_h(L_1, L_2) = \frac{2W_l(L_1, L_2)W_w(L_2)}{W_l(L_1, L_2) + W_w(L_2)}, \tag{12}$$

where we treat L_1 as the SWL and L_2 as the bridging LWL. We use the higher value of S_h to select better bridging LWLs.

[5] http://www.alphadictionary.com.
[6] https://www.wiktionary.org/.
[7] http://wals.info/.

2.4.2. Learning to select the best bridging language

The above approaches for ranking the bridging LWLs are handcrafted similarities. We also tried to use machine learning to learn from the features and generalize to other languages. Suppose we have a language pair L_i and L_j. We can construct a feature vector $\mathbf{x}_{ij}$ based on the WALS data, where the rth feature is:

$$\mathbf{x}_{ij}^{(r)} = I_r(L_i, L_j), \tag{13}$$

where the indicator function denotes both languages sharing the same WALS feature value.

If we consider L_i as the SWL, and there are two candidates LWLs L_j and L_k, we can compare L_j and L_k based on their feature vectors by projecting them to a real value: $\mathbf{w}^T\mathbf{x}_{ij}$ and $\mathbf{w}^T\mathbf{x}_{ik}$. Thus, if we have a lot of such pairs, we can build a support vector machine to learn the projection vector $\mathbf{w}$:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i\in\{SWL\},j,k\in\{LWL\}} \ell(\mathbf{w}^T\mathbf{x}_{ij} - \mathbf{w}^T\mathbf{x}_{ik}) \tag{14}$$

where ℓ is the loss function $\ell(t) = \max(0, 1-t)$ [20] and C is the penalty parameter. Then given the learnt $\mathbf{w}$, for any pair of LWLs L_j and L_k, we can evaluate which one is better to be used to bridge the SWL L_i based on $\mathbf{w}^T\mathbf{x}_{ij} - \mathbf{w}^T\mathbf{x}_{ik}$.

3. Experiments

In this section, we present our experiments on cross-lingual document classification with either large Wikipedia and small presence of Wikipedia data. Our experiments are designed to study the effectiveness of zero-shot classification in comparison to "standard" supervised classification algorithms, and to study the contribution of different semantic representations to the success of the zero-shot classification scheme.

3.1. The importance of representations

In this section, we first show how we build CLESA representations in many languages. We also present and compare an alternative way to do cross-lingual zero-shot classification: translating labels into each target language and then applying monolingual zero-shot classification in the target language. We note that this alternative shares some advantages with our proposed CLESA based method: it does not require heavy resources in the target languages (only the label space is to be translated) as do the other methods we mentioned earlier. However, when we compare this naive method with our proposed approach on a multilingual classification dataset, it turns out that our CLESA representation is the better choice for many language pairs while also being cheaper in terms of acquiring resources (no translation is needed). Note that we do not compare with another naive approach which translates both documents and labels to English and performs English ESA. This is because: (1) In practice, translation of documents is more costly than translation of labels and requires significantly more resources (label translation can be done once by an expert); (2) There is no large collection of documents in different languages labeled in the same label space to facilitate a fair comparison.

3.1.1. Building CLESA representations

We first downloaded the complete Wikipedia corpus (version available on August 5th, 2015) that is available in 180 languages including English.[8] The Wikipedia pages were tokenized and cleaned using the 38 available Lucene[9] language-dependent tokenizers[10] and with a whitespace based tokenizer for other languages. We filtered out pages with fewer than 100 words or 5 language links. This way, most of the redirection and disambiguation pages were removed and some of the short pages were also removed.

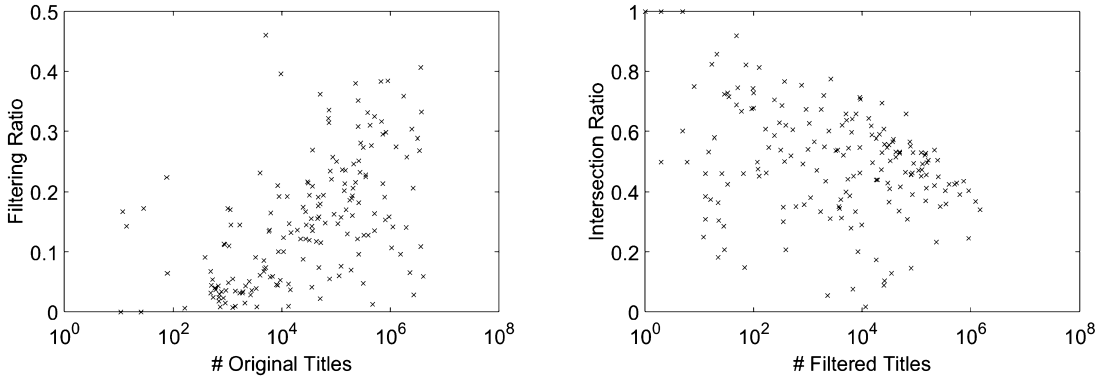
We fixed the English label space. Suppose that the target documents are in a foreign language L; in order to map the documents and labels to a common semantic space, we compute the intersection of the Wikipedia title pages linked between English and L. That is, for each language L, we only keep those Wikipedia pages that are linked to the English Wikipedia. This results in further reduction in the size of the collection available in each language.

Table 1 shows statistics about numbers of titles in the original 179 languages excluding English, after filtering and after intersection with English. There are 62 languages with more than 10,000 Wikipedia titles that are linked to the English Wikipedia. For these 62 languages we therefore have a title space that covers a wide range of topics. In Fig. 1(a) we show the ratio of remaining titles in each language after filtering short and non-linked pages (threshold = 5). This indicates that larger Wikipedias also tend to have longer and higher quality content. In Fig. 1(b), the ratio after intersecting with the English Wikipedia shows that larger-size Wikipedias have a more stable fraction of titles that are linked to the English Wikipedia, relative to smaller-size Wikipedias.

[8] https://dumps.wikimedia.org/.
[9] https://lucene.apache.org/.
[10] Stop words are embedded.

Table 1
Statistics of 179 Wikipedia corpora. "Filtered" corresponds to the numbers of pages after filtering with 100 words and 5 links. This is the data used for monolingual ESA. "$L \cap$ English" corresponds to the CLESA data.

# Titles	# Languages		
	Original	Filtered	$L \cap$ English
$n \geq 10^6$	17	2	0
$10^6 > n \geq 10^5$	50	27	14
$10^5 > n \geq 10^4$	45	44	48
$10^4 > n \geq 10^3$	39	41	41
$10^3 > n \geq 10^2$	21	25	31
$10^2 > n \geq 10$	7	31	24
$10 > n \geq 0$	0	9	21



(a) The ratio after filtering out Wikipedia pages with less than 100 words and 5 links.

(b) Intersection ratio of Wikipedia language links.

Fig. 1. The effects of preprocessing of Wikipedia. Each cross in the figures represents a language. (For English Wikipedia, originally we had around 15 million titles. After filtering, we had about 3 million titles.)

3.1.2. Monolingual ESA vs. CLESA

For cross-lingual document classification, a natural idea is to translate the documents and perform monolingual ESA. However, translation can be very costly and is not scalable to a large amount of documents. Another option is to keep the original set of Wikipedia titles in language L, and map the English label space to language L. This can be achieved with a relatively small effort compared with translating the documents, since the label space is rather small (i.e., no more than a few hundreds of words). Once we do that, we can generate an ESA representation in L, and run a monolingual zero-shot classification in L.

In order to understand the difference and relative advantages of this method and the one we proposed and presented earlier in Section 3.1.1, we perform the following experiment. We first translate a set of English documents to many languages, via Google Translate. (Note that we do this only to generate a new dataset on which we can perform a fair comparison of algorithms). We then perform the experiment as described above: translating the labels, and developing a monolingual ESA representation in language L, which is then used for zero-shot classification in language L. Specifically, we select 100 documents from the 20-newsgroups data set [42] which can be correctly classified using the English ESA. Then we use Google Translate API[11] to translate these documents into 88 languages.[12] We also filtered out Serbo-Croatian (sh) language since it has been deprecated and became a macro-language for Croatian (hr), Serbian (sr), Bosnian (bs) and Montenegrin (sr). We show the statistics of these 87 languages in Table 2. We also translate the 20 label descriptions to the 87 languages. We use the English label descriptions for the 20-newsgroups as in [24].

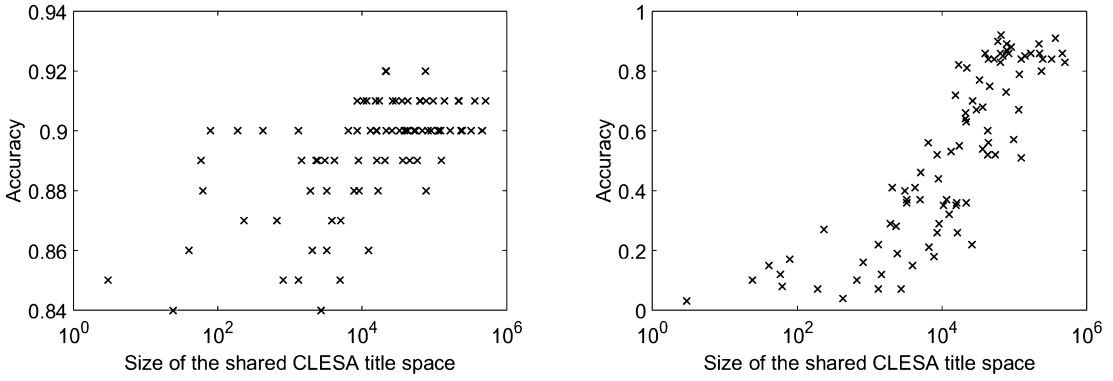
To test whether Google Translate will hurt the document quality, we perform the following evaluation. For the 100 documents in each language translated by Google, we translated them back to English again using Google Translate. Then we performed the English ESA based zero-shot classification [24]. A perfect translation should result in 100% accuracy. As shown in Fig. 2(a), The average classification accuracy is 0.893 ± 0.019, which seems good enough for us to use the translated documents as our evaluation data. The correlation score between the logarithm number of Wikipedia titles used

[11] https://github.com/mouuff/Google-Translate-API.

[12] Google translates only supports 88 out of the 179 languages that our CLESA method can deal with.

Table 2
Statistics of 87 Wikipedia corpora intersected with Google translation. "Filtered" corresponds to the numbers of pages after filtering with 100 words and 5 links. This is the data used for monolingual ESA. "$L \cap$ English" corresponds to the CLESA data.

# Titles	#Languages		
	Original	Filtered	$L \cap$ English
$n \geq 10^6$	16	2	0
$10^6 > n \geq 10^5$	43	25	14
$10^5 > n \geq 10^4$	20	34	40
$10^4 > n \geq 10^3$	5	19	22
$10^3 > n \geq 10^2$	3	3	5
$10^2 > n \geq 10$	0	4	5
$10 > n \geq 0$	0	0	1



(a) Zero-shot classification on documents translated back to English. Mean: 0.893; Std: 0.019.

(b) CLZSC.

Fig. 2. Correlation of Wikipedia page log-number and classification accuracy. Each cross represents a language. (a) $\rho = 0.604$, $p = 5.6 \times 10^{-10}$. (b) $\rho = 0.834$, $p = 1.1 \times 10^{-23}$. (ρ: Pearson's correlation coefficient. p: the significance value at level 0.05.)

in CLESA and the accuracy of translated English documents for 87 languages is $\rho = 0.604$ ($p = 5.6 \times 10^{-10}$). It seems that Google Translate's performance is also correlated with the size of acquired resource.

Now we can compare two settings for performing zero-shot classification: using monolingual ESA and using CLESA. We show the results of "top-1 label hit" and "top-3 labels hit" precisions in Table 3. Top-1 label means that for each document, we select the best label to classify it. This is exact classification evaluation. While for Top-3 labels, we select the best three labels for each document and check whether they contain the correct label. Comparing monolingual ESA and CLESA, the results in Table 3 show that even though the number of titles (Wikipedia pages) used by CLESA is much smaller than monolingual ESA, CLESA produces, on average, more accurate classifications. The language links used by CLESA help to disambiguate some Wikipedia titles. For example, some entities such as "python" have multiple meanings, which are better disambiguated when considering multiple languages. Therefore, the shared semantic space generated by CLESA provides a better representation than the single language title space. We also show the correlation between number of Wikipedia titles used in CLESA and the accuracy of CLZSC for 87 languages in Fig. 2(b). As shown in Fig. 2(b), the correlation score between the logarithm number of Wikipedia titles used in CLESA and the accuracy for 87 languages is $\rho = 0.834$ ($p = 1.1 \times 10^{-23}$). The classification result is significantly correlated with the logarithm number of Wikipedia titles.

3.1.3. *CLESA vs. cross-lingual embeddings*

We use the pretrained embeddings provided by [15][13] to test cross-lingual zero-shot classification. It uses FastText [43][14] to train monolingual embeddings and then learn the mappings between languages to align the word embeddings to be in the same space. The dimension of FastText embedding is 300. We call this set of embeddings as SVDAlign [15]. For the current version of SVDAlign, it contains 78 languages with cross-lingual embeddings, in which 66 languages can be used for our 20-newsgroups classification data. Fig. 3 shows the results. From the figure we can see that, CLESA outperforms SVDAlign on most of the languages overlapped between two approaches. The trends are also similar, which means better CLESA is correlated with embeddings for different languages.

[13] https://github.com/Babylonpartners/fastText_multilingual.
[14] https://github.com/facebookresearch/fastText.

Table 3
Precision statistics of 20-newsgroups classification in 87 languages. The numbers in the four columns represent the number of languages among the 87 for which the precision values fall within the ranges indicated on the left. Mono. stands for monolingual ESA. Cross. stands for cross-lingual ESA.

	Top-1 precision		Top-3 precision	
	Mono.	Cross.	Mono.	Cross.
$1 \geq p \geq 0.9$	2	20	5	28
$0.9 > p \geq 0.8$	7	8	9	8
$0.8 > p \geq 0.7$	10	7	16	8
$0.7 > p \geq 0.6$	14	4	13	8
$0.6 > p \geq 0.5$	8	9	8	5
$0.5 > p \geq 0.4$	6	5	4	7
$0.4 > p \geq 0.3$	10	9	10	6
$0.3 > p \geq 0.2$	6	9	6	5
$0.2 > p \geq 0.1$	7	8	5	4
$0.1 > p \geq 0$	17	8	11	8

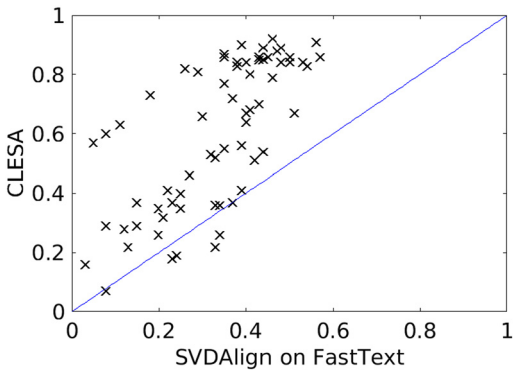


Fig. 3. Comparison of CLESA (Mean: 0.599; Std: 0.248) and SVDAlign on FastText (Mean: 0.332; Std: 0.136) with matched 66 languages. $\rho = 0.703$ and $p = 4.7 \times 10^{-11}$. (ρ: Pearson's correlation coefficient. p: the significance value at level 0.05.)

3.1.4. Extension with BabelNet

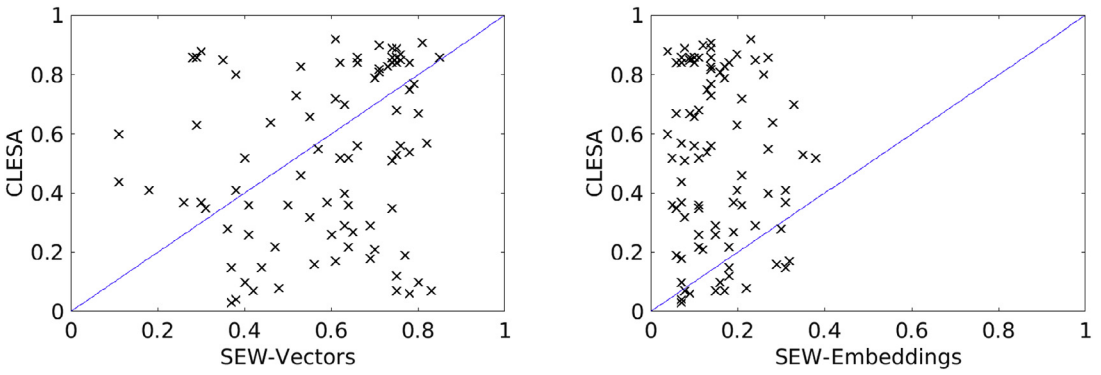
There could be many further improvements of CLESA. One way is to enrich the Wikipedia with more links [44] based on BabelNet [45], which contains not only Wikipedia knowledge but also rich semantic relationships provided by WordNet [46]. Since there are a lot of missing links, adding more links can both enrich the potential in-language disambiguation and language links between languages. As we have shown, there are still a lot of Wikipedias of small sizes. Moreover, using sense disambiguation may further improve the accuracy of CLESA results using Wikipedia. We also compare the semantically enriched Wikipedia (SEW) based multilingual representation based on BabelNet [47]. Since the vector representation is built based on BabelNet's synset, we first do word expansion to replace each word in a document with all its senses in BabelNet. In this case, we use the TF score of each synset to reweigh different senses in the document. After that, for each document, we use the top 500-concept as a vector representation, for which we call SEW-Vectors. The result comparing SEW-Vectors and CLESA is shown in Fig. 4(a). From the result we can see that SEW-Vectors are better than CLESA on average, as it uses enriched Wikipedia as resource to compute the representations. Interestingly, SEW-Vectors and CLESA are not correlated. This may be because the enrichment of Wikipedia and aggregation of word sense may significantly change the property of ESA representations. We also compare the semantically enriched Wikipedia based multilingual word embedding based on Word2vec [48,49]. We follow the same way of using BabelNet's synsets and use the averaged synsets embedding as the document representation. The result is shown in Fig. 4(b). The result is not as good as sparse vector representations. It may be because we haven't conducted sense disambiguation as many other applications used [48,50,51].

3.2. Cross-lingual classification on benchmark datasets

Section 3.1 established that the use of a common semantic space is a better way to perform zero-shot classification, and therefore we evaluate our proposed CLESA method in the standard document classification task.

3.2.1. Experimental settings

We present benchmark results for cross-lingual zero-shot classification on two datasets, TED and RCV2.



(a) Comparison of CLESA (Mean: 0.515; Std: 0.286) and SEW-Vectors (Mean: 0.590; Std: 0.185).

(b) Comparison of CLESA (Mean: 0.515; Std: 0.286) and SEW-Embeddings (Mean: 0.155; Std: 0.082).

Fig. 4. Evaluation of BabelNet. (a) $\rho = 0.184$, $p = 0.025$. (b) $\rho = -0.065$, $p = 2.0 \times 10^{-18}$. (ρ: Pearson's correlation coefficient. p: the significance value at level 0.05.)

TED Dataset. The TED dataset[15] is a multi-label classification dataset containing 15 labels of topics which are extracted from the most frequent keywords in the dataset. It is derived from a spoken language translation dataset.[16] It contains 13 languages from the TED talk transcriptions and their translations. The data has been organized into subsets according to each label where each subset contains about 1,200 documents (about 200 documents for "true" category for a label and 1,000 documents for "false" category).

RCV2 Dataset. RCV2 dataset is a multilingual extension of RCV1 [52]. We use the training set split by [52], however we do not stem the documents. Same as RCV1 data, RCV2 is newswire stories from Reuters Ltd under the Factiva news category taxonomies. There are different categorization methods in RCV2, e.g., topical or regional. For the topical categories, it contains 103 categories including all nodes except for root in the hierarchy. The maximum depth is four, and 82 nodes are leaves. Following cross-lingual document classification [6], we use the top level categories for evaluation. There are four categories which are GCAT (government social), ECAT (economics), MCAT (markets), and CCAT (corporate industrial). We aggregate all the subtree's English descriptions for each category as the category description. RCV2 data contains documents in 13 languages.

Since TED dataset is a multi-label classification dataset, we report averaged F1 over 15 labels, which results in macro-F1 score. For RCV2 dataset, we have processed it to be a multi-class classification problem by removing documents with multiple labels. However, different categories have different numbers of documents. Thus, here we use both macro-F1 and micro-F1 scores to evaluate the results instead of just using the accuracy. Both macro-F1 and micro-F1 scores have been widely used in text classification [53].

Before discussing the results, we first list the baselines and our comparison methodology. Our goal is to evaluate the quality of classifying documents in Language L into an English label space. To understand the advantages and shortcomings of our method, we compare the following approaches.

CLESA and bootstrapping. We use the CLESA described in Section 2.3 for zero-shot classification for both datasets. We also use bootstrapping described in Section 2.2 to further enhance the unsupervised learning results.

Supervised learning. To compare how good zero-shot classification can be, we implemented supervised baselines for both datasets. We use simple BOW representation of documents (tokenized with stop words removed by Lucene), and use Liblinear [54] as the classifier. Particularly, we use the L2-regularized and L2-loss linear support vector classification for all the experiments.

Cross-lingual word embedding. Another approach to perform cross-lingual zero-shot classification is to embed the words in both languages into the same semantic space, and then compare documents and labels in different languages in the same space. We first use the compositional vector model (CVM) [37] to generate our bi-lingual word embedding in a shared semantic space. CVM needs parallel corpora to train embedding for both languages. Following [37], we train the models based on TED and Europarl[17] datasets. Europarl dataset is a popularly used parallel corpus for machine translation [55]. It has 21 European languages that can be translated into English and vice versa. In this experiment, we select the ten relevant languages to the cross-lingual zero-shot classification tasks.

The statistics of both TED and Europarl datasets is shown in Table 4. In the table, we also show our comparable corpus provided by Wikipedia language links. From the table we can see that Europarl data is about an order of magnitude larger than TED data. Wikipedia is even larger in terms of token numbers (except for Danish and Swedish). But Wikipedia is nosier

[15] http://www.clg.ox.ac.uk/tedcldc.html.
[16] https://wit3.fbk.eu/.
[17] http://www.statmt.org/europarl/.

Table 4
Statistics of parallel (TED and Europarl) and comparable (Wikipedia) corpora. "Lang" denotes the statistics for the foreign language while "English" means the statistics in English in a parallel or a comparable corpus. "Tok." denotes the number of tokens (in millions) in the corpora and "Voc." denotes the vocabulary size (in thousands).

	TED				Europarl					Wikipedia				
	Lang		EN		Lang		EN			Docs	Lang		EN	
	Tok.	Voc.	Tok.	Voc.	Tok.	Voc.	Tok.	Voc.			Tok.	Voc.	Tok.	Voc.
Arabic	2.29	73.8	2.95	43.9	–	–	–	–	Arabic	78K	39	1,000	120	1,910
Danish	–	–	–	–	44.7	647	48.7	307	Danish	64K	20	850	104	1,710
German	2.60	45.8	2.75	42.7	44.6	649	47.9	305	German	504K	236	5,060	403	5,040
Spanish	2.87	42.5	3.06	44.7	51.6	423	49.2	309	Spanish	368K	172	2,980	344	4,350
French	3.13	40.8	3.08	44.4	52.5	418	50.3	312	French	456K	217	2,830	389	4,740
Italian	2.81	45.7	3.09	44.9	48.0	455	49.7	309	Italian	327K	158	2,490	307	3,920
Japanese	–	–	–	–	–	–	–	–	Japanese	97K	145	1,580	142	2,170
Dutch	2.58	38.9	2.82	43.0	50.7	524	49.5	309	Dutch	235K	73	1,990	237	3,310
Norwegian	–	–	–	–	–	–	–	–	Norwegian	124K	38	1,380	158	2,400
Portuguese	2.85	39.8	3.02	44.5	50.0	443	49.3	310	Portuguese	218K	84	1,610	245	3,260
Polish	2.20	68.3	2.88	43.5	12.8	339	15.3	149	Polish	246K	90	1,870	249	3,350
Romanian	2.94	53.0	3.07	44.8	9.6	178	9.7	114	Romanian	53K	28	830	90	1,600
Russian	2.21	63.7	2.57	41.3	–	–	–	–	Russian	222K	121	2,490	231	3,460
Swedish	–	–	–	–	41.6	622	45.8	298	Swedish	168K	47	1,690	205	2,910
Turkish	1.90	71.9	2.63	41.7	–	–	–	–	Turkish	63K	26	910	106	1,740
Chinese	0.68	13.2	2.99	44.2	–	–	–	–	Chinese	43K	97	600	85	1,520

and the distribution of words is imbalanced, e.g., English tokens can be five times larger compared to Danish and four times larger compared to Swedish. This means in general, English Wikipedia pages are longer than other languages.

We trained the CVM model using the parallel corpora with the default setting as well as the settings indicated in the paper [37] using their software.[18] The length of the word vector was set to 128, the number of iterations was set to five, and the number of mini-batches was set to ten. We used the "additive" model with single mode (only using pairwise languages) and used the "doctrain" model to train on TED data and the "dbltrain" model to train on Europarl data.

We also use the pretrained embeddings provided by [15] (SVDAlign) and MUSE embeddings [16].[19] MUSE embeddings are also developed based on FastText [43]. Different from SVDAlign, MUSE does not require any bilingual lexicon to train the cross-lingual embeddings. For the current version of MUSE, it contains 30 languages with cross-lingual embeddings (while as mentioned SVDAlign released 78 languages).

3.2.2. TED data classification

The TED data has already been organized into subsets according to each label. Thus, we treat the problem as a binary classification for each label. Since the data is imbalanced, we find that training a supervised binary classifier and using the default threshold to determine which one is positive is not effective enough. Thus, we randomly split the provided training set into 70% training and 30% validation sets. Then we use the training set to train a model and use the validation set to tune the threshold. We average the results over ten trials to select the best threshold. Then we train a new model using the full training data and apply the new model and the tuned threshold to the test set. Besides using the full training set, we also use 10% and 15% of the full training set to do the same supervised procedure, respectively. We report the averaged F1 scores over 10 trials for 15 labels with supervised learning in Table 5. The fully supervised learning results are comparable with the best results shown in Table 4 in [37].

For zero-shot classification, since there is only one label for each binary classification problem, it is only possible to use one similarity to select the most similar documents and label them as positive. Therefore, we perform a naive zero-shot classification as follows. First, we merge the training and testing datasets, which contains around 1200 documents for each label and each language. Then we select the 200 highest similarities between each label and the documents, and label them as positive. For bootstrapping, we initially label 50 positive and 500 negative examples respectively, and train a classifier, and then iteratively label 5 positive and 50 negative more documents in each bootstrapping step. We also combine the bootstrapping results with the top 200 positive documents labeled with pure zero-shot classification to ensure good recall. In addition, we also use another setting to verify the cross-lingual ESA similarity. We use the training set to tune a threshold for the similarities computed by cross-lingual ESA between both labels and documents. Then we apply the threshold to the test set to classify the documents. We call this the "tuned zero-shot classification."

From Table 5 we can see that naive zero-shot classification with the top 200 documents performs worst among the three settings, while bootstrapping is in the middle and tuned zero-shot classification performs the best. Compared to supervised learning, zero-shot classification is comparable to supervised learning with 10% labeled data, and a little worse than supervised learning with 15% labeled data. This result is consistent with the results shown in the original monolingual

[18] https://github.com/karlmoritz/CVM.
[19] https://github.com/facebookresearch/MUSE.

Table 5
Comparison on TED dataset (averaged macro-F1 scores over 15 labels). CLESA naive: merging training and test data, selecting the top 200 highest similarity scores as positive, and evaluating the F1 score on test set. CLESA bootstrap: bootstrapping over the naive method. CLESA tuned: tuning a threshold on the training set, and applying it on the test data. Word2vec (skipgram modeled trained on Wikipedia with 128 dimensions) on English is 0.346 with tuned setting. "Average" excludes English.

	Supervised			CLESA			Embedding (tuned)			
	Full	10%	15%	Naive	Bootstrap	Tuned	TED	Europarl	SVDAlign	MUSE
English	0.508	0.316	0.360	0.389	0.405	0.440	–	–	0.373	0.377
Arabic	0.468	0.223	0.286	0.273	0.299	0.266	0.240	–	0.350	0.323
German	0.449	0.234	0.278	0.222	0.245	0.248	0.219	0.115	0.341	0.329
Spanish	0.525	0.303	0.331	0.289	0.301	0.293	0.245	0.163	0.366	0.352
French	0.547	0.353	0.324	0.205	0.228	0.206	0.253	0.157	0.392	0.375
Italian	0.535	0.294	0.315	0.191	0.197	0.226	0.289	0.177	0.383	0.365
Dutch	0.494	0.308	0.319	0.340	0.360	0.390	0.285	0.157	0.364	0.368
Polish	0.420	0.209	0.296	0.227	0.253	0.286	0.278	0.174	0.369	0.342
Pt-Br	0.502	0.271	0.296	0.307	0.331	0.287	0.250	0.171	–	
Roman.	0.491	0.295	0.257	0.170	0.194	0.241	0.232	0.213	0.370	0.351
Russian	0.475	0.216	0.278	0.199	0.195	0.205	0.127	–	0.354	0.340
Turkish	0.426	0.176	0.252	0.333	0.354	0.395	0.248	–	0.329	0.339
Chinese	0.235	0.158	0.167	0.173	0.182	0.239	0.197	–	0.232	
Average	0.468	0.258	0.289	0.255	0.273	0.286	0.238	0.166	0.350	0.348

zero-shot classification [23,24]. It is amazing that for Chinese document classification, zero-shot classification is even better than the fully supervised learning. This may be because that Chinese typically uses fewer segmented words than English to represent the same meanings (0.68 million tokens in Chinese vs. 2.99 million tokens in English in TED). Then when classification is conducted on BOW features, there are fewer overlapped words among documents in Chinese, as compared to English and other languages.

We use the multilingual embeddings for the zero-shot classification setting. We only show the results based on the tuned classification approach (tuning threshold based on training and applying the threshold for testing) in Table 5. Since Europarl data cannot cover all the language pairs used in TED, we only report the ones that it can cover. From the results we can see that even though the Europarl dataset is much larger than TED, the embedding results trained based on TED data are much better than the embedding trained based on the Europarl dataset. The results based on SVDAlign and MUSE are better than CVM embeddings and CLESA. This is because, for the comparison with CVM, SVDAlign and MUSE are both based on larger training corpus and more advanced algorithms. For the comparison with CLESA, since the TED data only use one label keyword to describe the categories, the dense embedding can better compress the meanings of labels and compare with the texts. On the contrary, CLESA may be more difficult to find a good representation based on single label description. As we can see for the 20-newsgroups data and RCV2 data, CLESA is still better than SVDAlign and MUSE since the labels to describe the categories are better used.

We also verified embedding results by testing English language with word2vec [29,30] trained with Skipgram model, vector length as 128, and window size as five on the whole English Wikipedia. The tuned zero-shot classification result is 0.346, which is less than the English ESA (0.440) shown in Table 5. Similar results have also been shown in previous monolingual zero-shot classification [24].

3.2.3. RCV2 data classification

We use the linear classifier trained on BOW as the baseline method. We train the classifiers with 400 and 800 randomly selected examples for each language respectively. We report the average over 10 trials for supervised learning results. For zero-shot classification, we have four classes and we choose the best label for each document based on the highest similarity between a document and the label descriptions. Then for bootstrapping, we use the standard procedure to initialize 100 documents for each class using pure similarity based zero-shot classification, and then iteratively label 100 more documents for each class and stop after three iterations. From Tables 6 and 7 we can see that zero-shot classification with bootstrapping is comparable to supervised learning using between 400 and 800 labeled documents.

For the zero-shot classification based on multilingual embedding, we can see that the embedding trained on TED performs worse than embedding on Europarl. Compared to the TED classification results where TED embedding is much better, now both TED embedding and Europarl embedding are applied to out-of-domain examples (RCV2 words). Thus, when changing the domain, the size of the training corpus matters. Again, SVDAlign and MUSE are better than CVM for the same reason as explained for TED data. Cross-lingual ESA is better than cross-lingual embeddings. This is reasonable, since for embedding, we average all the word vectors to represent a document and a label. Thus some information may be lost.

In addition, the zero-shot classification with word2vec embedding [30] with 128 dimensions trained on English Wikipedia for RCV1 is 0.561. This again verifies that embedding currently under-performs ESA for zero-shot classification.

Table 6
Comparison on RCV1/RCV2 datasets (top level, four categories) on micro-F1. S.400: supervised learning with 400 training data. S.800: supervised learning with 800 training data. CLESA: zero-shot classification with CLESA. Bootstrap: zero-shot classification with CLESA+bootstrapping. E.(TED): word embedding using CVM trained on TED data, document embedding with average word embedding. E.(Euro.): word embedding using CVM trained on Europarl data. "Average" excludes English. "w/o zh/ja" means average without considering Chinese and Japanese. Zero-shot classification with word2vec embedding [30] with 128 dimensions trained on English Wikipedia for RCV1 is 0.561.

micro-F1		Supervised		CLESA		Embedding			
	#Doc.	S.400	S.800	CLESA	Bootstrap	E.(TED)	E.(Euro)	SVDAlign	MUSE
RCV1	23,149	0.691	0.786	0.653	0.742	–	–	0.615	0.619
Danish	11,185	0.589	0.630	0.317	0.364	–	0.352	0.456	0.485
German	116,212	0.424	0.492	0.613	0.724	0.396	0.305	0.629	0.621
Spanish	18,655	0.645	0.651	0.647	0.667	0.156	0.290	0.162	0.166
Sp.-latam	79,775	0.241	0.250	0.644	0.554	0.376	0.536	0.722	0.722
French	85,393	0.307	0.467	0.653	0.762	0.578	0.334	0.633	0.671
Italian	28,406	0.553	0.607	0.528	0.542	0.323	0.274	0.346	0.322
Japanese	65,499	0.548	0.595	0.324	0.534	–	–	0.114	–
Dutch	1,794	0.140	0.160	0.387	0.395	0.125	0.205	0.754	0.717
Norwegian	9,409	0.510	0.564	0.252	0.329	–	–	0.554	0.542
Portuguese	8,841	0.546	0.613	0.428	0.375	0.101	0.257	0.152	0.129
Russian	17,487	0.499	0.523	0.309	0.418	0.334	0.323	0.155	0.153
Swedish	15,732	0.454	0.518	0.466	0.618	–	0.330	0.497	0.491
Chinese	28,964	0.672	0.723	0.537	0.690	0.241	–	0.094	–
Average		0.471	0.523	0.470	0.536	0.292	0.320	0.405 (0.460 w/o zh/ja)	0.456

Table 7
Comparison on RCV1/RCV2 datasets (top level, four categories) on macro-F1. S.400: supervised learning with 400 training data. S.800: supervised learning with 800 training data. CLESA: zero-shot classification with CLESA. Bootstrap: zero-shot classification with CLESA+bootstrapping. E.(TED): word embedding using CVM trained on TED data, document embedding with average word embedding. E.(Euro.): word embedding using CVM trained on Europarl data. "Average" excludes English. "w/o zh/ja" means average without considering Chinese and Japanese. Zero-shot classification with word2vec embedding [30] with 128 dimensions trained on English Wikipedia for RCV1 is 0.465.

macro-F1		Supervised		CLESA		Embedding			
	#Doc.	S.400	S.800	CLESA	Bootstrap	E.(TED)	E.(Euro)	SVDAlign	MUSE
RCV1	23,149	0.747	0.764	0.586	0.698	–	–	0.474	0.465
Danish	11,185	0.454	0.480	0.322	0.389	–	0.271	0.344	0.380
German	116,212	0.388	0.472	0.564	0.685	0.276	0.263	0.414	0.434
Spanish	18,655	0.321	0.343	0.543	0.625	0.151	0.181	0.147	0.157
Sp.-latam	79,775	0.343	0.381	0.494	0.525	0.204	0.273	0.288	0.297
French	85,393	0.395	0.525	0.583	0.662	0.230	0.246	0.283	0.340
Italian	28,406	0.423	0.480	0.520	0.552	0.323	0.274	0.267	0.258
Japanese	65,499	0.410	0.448	0.321	0.517	–	–	0.054	–
Dutch	1,794	0.212	0.227	0.316	0.344	0.125	0.205	0.350	0.329
Norwegian	9,409	0.407	0.427	0.229	0.309	–	–	0.271	0.282
Portuguese	8,841	0.305	0.403	0.389	0.361	0.101	0.257	0.142	0.125
Russian	17,487	0.336	0.382	0.292	0.366	0.334	0.323	0.089	0.083
Swedish	15,732	0.389	0.429	0.444	0.580	–	0.330	0.327	0.329
Chinese	28,964	0.412	0.507	0.465	0.603	0.241	–	0.047	–
Average		0.369	0.423	0.422	0.501	0.221	0.262	0.233 (0.266 w/o zh/ja)	0.274

3.3. Cross-lingual document classification with small Wikipedia

In this section, we present the experimental results of document classification when there is little or no Wikipedia presence. CLZSC can achieve relatively good performance when the number of Wikipedia titles is large. Since the correlation between CLZSC results and Wikipedia sizes is significant, we split the 87 languages based on the classification results. There are 39 languages with lower than 0.5 classification accuracy, while 48 with higher than 0.5 accuracy. In the following experiments, we call the 39 languages the SWLs and the 48 languages as well as English (in total 49) the LWLs.

3.3.1. Example languages: Hausa and Uzbek

We first select two typical SWLs, i.e., Hausa and Uzbek, as examples to demonstrate how to use bridging languages to improve zero-shot classification. Hausa is a language under the Afro–Asiatic family and further under Chad. Uzbek is a language under the Middle Turkic family. Both of the writing systems are related to Latin. The small number of Wikipedia pages, and therefore small shared semantic space, for these two languages (62 for Hausa and 3,082 for Uzbek after intersection with English Wikipedia), means that CLZSC will not be accurate. The results are summarized in Table 8. Indeed, classification results are not satisfactory (0.08 for Hausa and 0.40 for Uzbek).

Table 8
Comparison on Hausa and Uzbek languages data. "en" stands for English. "ar" stands for Arabic. "ha" stands for Hausa. "uz" stands for Uzbek. "ACC." stands for accuracy. "Pur." stands for purity.

ZERO-SHOT CLASSIFICATION	LG=HA (ACC.)	LG=UZ (ACC.)
lg-en Wiki.	0.08	0.40
en Wiki. (lg-en dict.)	0.27	0.63
ar-en Wiki. (lg-en dict.)	0.43	0.71
ar-en Wiki. (lg-ar dict.)	0.75	0.79
CLUSTERING	LG=HA (PUR.)	LG=UZ (PUR.)
K-means (en)	0.714 ± 0.025	–
K-means (ar)	0.698 ± 0.046	–
K-means (lg)	0.684 ± 0.025	0.686 ± 0.031

The basic idea of bridged CLZSC is that if we can leverage some word level translation from SWLs to another language, we can use the other language to build ESA/CLESA and further perform zero-shot classification. Here we tried to use both English (3 million titles) and Arabic (77,631 intersected titles) to bridge Hausa and Uzbek. To compare dictionaries, we first used Google Translate to translate all the words (word by word) used in 20-newsgroups documents in Hausa and Uzbek to English. Then the zero-shot classification result using English ESA is 0.27 and 0.63 for Hausa and Uzbek respectively.

To test the CLZSC using Arabic as a bridging language, we use Google Translate to translate Hausa/Uzbek words into Arabic words. Then we map each document in Hausa/Uzbek to Arabic, and perform CLESA based on Arabic–English Wikipedia. The result of zero-shot classification is 0.75 and 0.79 for Hausa and Uzbek respectively. We presume that there are two potential reasons for Arabic being better than English. First, the Arabic–English intersected space may be less ambiguous than the original English space. The language links used by CLESA reduce the size of space of Wikipedia titles, but help to disambiguate the semantic meanings. Second, the word-to-word mappings for Hausa/Uzbek–Arabic are better than those for Hausa/Uzbek–English because Hausa/Uzbek and Arabic are in the same writing system. To test the two above hypotheses, we also map Hausa/Uzbek to English and use the English part in the Arabic–English intersected Wikipedia to perform zero-shot classification. The results are 0.43 and 0.71 respectively, less than using the Arabic part but greater than using Hausa/Uzbek–English mapping for English Wikipedia. We summarize all the above results in Table 8.

We also compare CLZSC with the traditional unsupervised clustering algorithm, K-means, over the TF-IDF features of documents (IDF was computed based each language's 100 documents). Note that in zero-shot classification, we only need label names to classify the documents, but K-means needs to know a set of documents in the target language. When seeing more documents, CLZSC can also be further improved by bootstrapping. We performed ten trials and average the results, using the purity metric[20] to evaluate the accuracy of clustering. Purity is an average accuracy of each cluster assigned to the max corresponding ground truth label. It can be regarded as an upper-bound accuracy when we do not know the correspondence between the ground truth labels and the clustered labels. The clustering results are comparable for English, Arabic, and Hausa, but not as good as bridged CLZSC. In addition, from the results we can see that there is no clear clue about which language will have better clustering results.

3.3.2. Bridged cross-lingual zero-shot classification

Given the fact that for both Hausa and Uzbek, Arabic outperforms English for bridged CLZSC, and the fact that there are more local languages out of 7,000 languages in the world that cannot be translated to English but may be able to be translated to local popular language, we want to evaluate which language can be the best language as a bridge for the SWLs. In Table 9, we show the top ten bridging languages for the target languages Hausa and Uzbek. All the translation of words are performed by Google Translate.[21] The results show that Arabic is the best bridge for both languages.

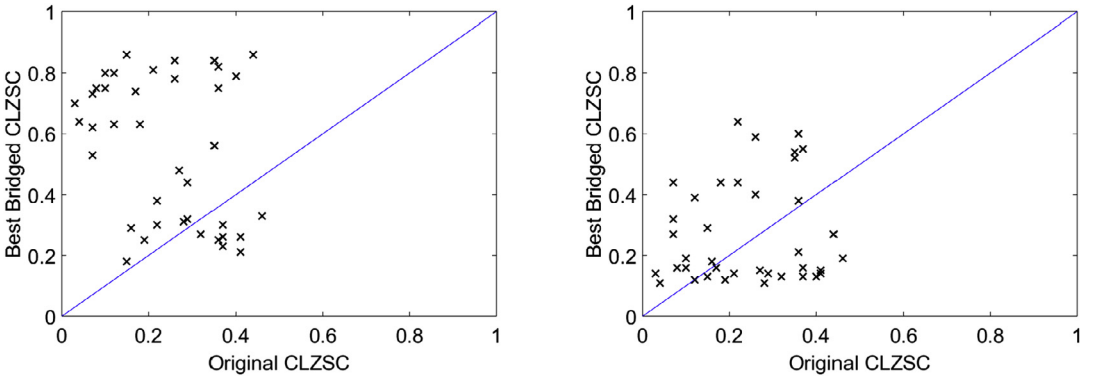
For all the 39 SWLs, we also checked the bridged results based on all the 49 LWLs, and we selected the best bridging languages and report the classification results. We compare the original CLZSC results based on the SWL-English Wikipedia with the best LWL bridged CLZSC in Fig. 5. We show the results using Google Translate in Fig. 5(a). We also show the results using Panlex word translation in Fig. 5(b). There are 8 out of 39 languages bridged CLZSC being worse than the original CLZSC using Google Translate, while there are 17 languages worse with Panlex translation. To further evaluate the quality of Panlex dictionaries, we traversed all the 6,134 distinctive language codes in Panlex. We found there are 1,671 languages with at least one word in the selected 100 documents in 20-newsgroups data that can be translated into English. The percentage of words that can be translated versus the number of expressions shown in Panlex is shown in Fig. 6. It turns out that only 12.39% out of 1,671 languages has more than 10% words identified. This is why Panlex translation results are worse than Google Translate shown in Fig. 5.

[20] https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html.

[21] Here we use Google Translate since it has consistent coverage across the evaluation data we used. Wiktionary (and other dictionaries) covers 1,000+ languages and makes our method a lot more scalable, but we have yet to systematically compare the quality of various dictionaries.

Table 9
Bridging languages ranks for Hausa and Uzbek translated 20-newsgroups data.

	Hausa		Uzbek	
Rk.	Bridge	Acc.	Bridge	Acc.
1	Arabic	0.75	Arabic	0.79
2	Hebrew	0.72	Korean	0.78
3	Korean	0.72	Hebrew	0.77
4	Bulgarian	0.67	Catalan	0.76
5	Persian	0.67	Bulgarian	0.76
6	Russian	0.67	Russian	0.71
7	Indonesian	0.65	Indonesian	0.70
8	Thai	0.65	Persian	0.70
9	Japanese	0.63	Spanish	0.70
10	Lithuanian	0.61	Japanese	0.70



(a) Bridged CLZSC with Google word translation. (b) Bridged CLZSC with Panlex word translation.

Fig. 5. Comparison of original CLZSC and best bridged CLZSC on 39 SWLs.

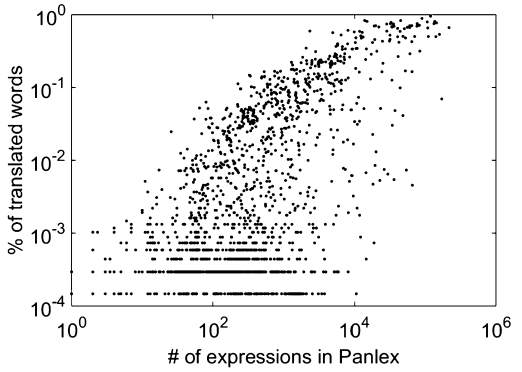


Fig. 6. Panlex word translation on 20-newsgroups data (1,671 languages to English).

3.3.3. Ranking based results

Table 10 shows the ranking based results.

"Original" row is the result for original CLZSC. We compute the mean and standard deviation for the 39 SWLs as well as the correlation of CLZSC with the best bridged CLZSC. The correlation value is negative. According to Fig. 5(a), it seems the improvement over smaller original CLZSC accuracies is larger than the ones with larger CLZSC accuracies.

"Majority Voting" row shows the results of using all the LWLs to vote for each zero-shot classification result. It shows that "Majority Voting" is significantly better than original CLZSC, and highly correlated with "Best Bridge" shown in next row.

"Best Bridge (Google)" row shows the results of bridged CLZSC results with the best bridge LWLs. This is the upper-bound of all the other ranking based methods. We can also see from Fig. 5(a) that the result is significantly better. Although the variance of the results is large, the t-test result still shows significance. "Best Bridge (Panlex)" shows no significant

Table 10
Comparison of different methods to select LWLs to bridge SWL CLZSC. The t-tests are performed between each ranking results with the original CLZSC results. The t-test p-value between "RankSVM" and "Majority Voting" is 0.014. The dependent correlation tests are performed with each value with the previous one(s), i.e., "Wikipedia size vs. Linguistic," "Wikipedia language links vs. Linguistic," "Combination vs. Linguistic (Ling.)," "Combination vs. Wikipedia size (Wiki.)," and "RankSVM vs. Combination (Comb.)." All the p-values are at 0.05 level (greater than 0.05 will reject the hypothesis).

METHOD	MEAN ± STD	T-TEST p-VALUE	CORR. W. BEST (GOOGLE)	DEPENDENT CORR. p-VALUE
Original	0.242 ± 0.126	–	−0.310	–
Majority Voting	0.438 ± 0.236	2.061×10^{-4}	0.974	
Best Bridge (Panlex)	0.269 ± 0.167	0.406	–	–
Best Bridge (Google), upper bound	0.546 ± 0.238	2.272×10^{-7}	1.000	–
Linguistic	0.380 ± 0.243	0.011	0.827	–
Wikipedia language links	0.275 ± 0.186	0.389	0.805	0.734 (Ling.)
Wikipedia size	0.277 ± 0.186	0.366	0.856	0.607 (Ling.)
Combination (wiki size)	0.353 ± 0.221	0.013	0.773	0.402 (Ling.), 0.096 (Wiki.)
RankSVM	0.465 ± 0.229	2.067×10^{-5}	0.963	3.646×10^{-5} (Ling.)

improvement over original CLZSC. However, Panlex has much more languages than Wikipedia and Google Translate. Thus, it might be still useful when there is something than no resource at all.

"Linguistic" row shows the results of bridged LWLs ranked by $S_h(L_1, L_2)$ in Eq. (9). It is significantly better than original CLZSC at 0.05 level.

"Wikipedia language links" row shows the results ranked by $S_{ll}(L)$ in Eq. (10). $S_{ll}(L)$ is almost the same as $S_w(L)$, since for most of the languages, English has the largest language link number. "Wikipedia size" row shows the results of bridged LWLs ranked by $S_w(L)$ in Eq. (10). $S_w(L)$ will always rank English as the bridge language. We have two interesting findings from the results. First, the ranking is not significantly better than original CLZSC, and worse than "Linguistic." This means that, bridging SWLs with English by mapping only words may not be a better solution compared to using cross-lingual Wikipedia, even though the cross-lingual Wikipedia is not good enough. Second, the correlation value between "Wikipedia size" and "Best Bridge" is higher than the correlation value between "Linguistic" and "Best Bridge." However, the dependent correlation test [56][22] shows this improvement is not significant.

"Combination" row shows the results of bridged LWLs ranked by $S_w(L)$ in Eq. (12). The results show that combining the "Linguistic" and "Wikipedia size" features by hand shows no improvement over pure "Linguistic" features.

"RankSVM" row shows the results using RankSVM. We split the SWLs into five folds. Then we perform a five-fold cross validation to generate the results. For each validation, we use 80% of the SWLs as training data, where each language has 49 LWLs accuracies. We use the 49 accuracies to generate $49 \times 48/2$ pairs. Then we use the learnt model to rank the other 20% SWLs. After the five-fold cross validation, we can rank all the SWLs based on each learnt model. We tune the parameter of C using a grid search in $\{10^{-2}, 10^{-1}, \ldots, 10^{4}\}$. The average result over 39 SWLs is significantly better than original CLZSC ($p = 2.067 \times 10^{-5}$) and "Majority Voting" ($p = 0.014$). The correlation with "Best Bridge" is also significantly better than "Linguistic" with "Best Bridge." This means that machine learning based method is significantly better than the unsupervised voting and ranking with handcrafted similarities. By looking into the averaged weights of RankSVM in five fold cross validation, we select the five top weights with largest absolute values, which are: "Internally-headed relative clauses" (−0.3613), "Front Rounded Vowels" (0.1656), "Absence of Common Consonants" (0.1538), "Optional Double Negation in SVO languages" (−0.1402), "Number of Genders" (0.1385).

4. Related work

In this section, we briefly survey some related work.

4.1. Cross-lingual classification

Cross-lingual document classification has attracted more attention recently in low-resource settings, where target language training data is minimal or unavailable. It is a natural sub-topic of transfer learning [57]. In cross-lingual document classification, we train a classifier on labeled documents in the *source* language, and classify documents in the *target* language. Existing approaches either need a parallel corpus to train word embeddings for different languages [37], require labeled documents in both source and target languages [36], make use of machine translation techniques to translate words [58] or documents [5], or combine different approaches [59]. Among the existing approaches, word translation is the cheapest way, while document translation and annotation on the target domain are the most expensive. In the middle, parallel or comparable corpora may be used to learn a good word/document representation, which avoids document translation but can still find a correspondence between source and target languages. The strength of cross-lingual document classification is that it can be generalized to multiple languages even in the absence of resources. However, when

[22] We used the implementation here: https://github.com/psinger/CorrelationStats.

we change the label space from one domain to another, we should perform translation again and re-train the classifiers. Instead of using parallel corpus to train a classifier or train a translation model, our approach only needs the comparable corpus, Wikipedia, in different languages aligned with English. Then if a user can tell the name of the category, cross-lingual zero-shot classification can perform text classification on-the-fly.

4.2. Pivot based machine translation

Pivot language is used to help machine translation when there is no enough resources to train a translation model from source language to target language [18,60–64]. For example, Paul et al. [18] used 22 Indo-European and Asian languages to evaluate how to select a good pivot language for machine translation. They evaluated 45 features falling into eight categories. Besides the language family feature, they used more translation-relevant features such as length of sentence, reordering, overlap of vocabulary, etc. They showed that the final result is mostly affected by the source-pivot and pivot-target translation performance. They also mentioned machine learning based method in the future work, but we are unaware of a follow-up paper that succeeded in doing it. Different from machine translation which needs the sentence level source-pivot and pivot-target translation, in cross-lingual classification, it is sufficient to use word dictionaries, making borrowing a bridging language more scalable to many languages and thus more practically useful. To the best of our knowledge, we have studied largest number of LWLs (49) and SWLs (39) with largest number of linguistic features (196).

4.3. Zero/one-shot learning

Zero-shot learning [7–10] and one-shot learning [11,12] were first introduced in the computer vision community and are now recognized by the natural language processing community [65,66]. One-shot learning requires one example for training, while in zero-shot learning, the test data is different from the training data (e.g., a new label space). However, in contrast to the zero-shot classification scenario, both learning protocols require some training data. The zero-shot classification protocol, on the other hand, assumes no direct training data but only the label names or descriptions. Compared to one-shot learning, the labels can be relatively simpler. In addition, it relies on background data from external knowledge sources (like Wikipedia), that is used in an unsupervised way to generate a common semantic space.

5. Conclusions

In this paper, we proposed a cross-lingual zero-shot classification, CLZSC, approach to text categorization. We show that it is possible to classify documents in multiple languages into an English label space without any training data. This framework maps target documents and label space into a shared cross-lingual text representations. Then we studied the problem of CLZSC for SWLs. CLZSC uses English labels to classify documents in other languages and is scalable to many languages and adaptive to any label space. However, if Wikipedia for a language is not large enough, the performance is not acceptable. We simply map the words in SWL documents to LWL words, and perform zero-shot classification based on LWLs. The experiments conducted on 88 languages derived from 20-newsgroups data show that for 28 languages, the pure zero-shot classification can achieve greater than 0.8 accuracy. We also tested on two multilingual benchmark datasets, i.e., TED and RCV2 datasets, showing that zero-shot classification is comparable to supervised learning with about 100 labeled documents per label. Finally, we systematically evaluate 39 SWLs and 49 LWLs. Experiments show that bridging the SWLs with LWLs can significantly improve the classification results.

Acknowledgements

The authors wish to thank the anonymous reviewers of conference papers. This work was supported by DARPA under agreement numbers HR0011-15-2-0025 and FA8750-13-2-0008. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the organizations that supported the work. Yangqiu Song is also supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 26206717).

References

[1] I. Dagan, Y. Karov, D. Roth, Mistake-driven learning in text categorization, in: EMNLP, 1997, pp. 55–63.
[2] T. Joachims, Text categorization with support vector machines: learning with many relevant features, in: ECML, 1998, pp. 137–142.
[3] S. Dumais, H. Chen, Hierarchical classification of web content, in: SIGIR, 2000, pp. 256–263.
[4] R. Agrawal, A. Gupta, Y. Prabhu, M. Varma, Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages, in: WWW, 2013, pp. 13–24.
[5] M. Amini, C. Goutte, A co-classification approach to learning from multilingual corpora, Mach. Learn. 79 (1–2) (2010) 105–121.
[6] A. Klementiev, I. Titov, B. Bhattarai, Inducing crosslingual distributed representations of words, in: COLING, 2012, pp. 1459–1474.
[7] M. Palatucci, D. Pomerleau, G.E. Hinton, T.M. Mitchell, Zero-shot learning with semantic output codes, in: NIPS, 2009, pp. 1410–1418.
[8] R. Socher, M. Ganjoo, C.D. Manning, A.Y. Ng, Zero-shot learning through cross-modal transfer, in: NIPS, 2013, pp. 935–943.
[9] M. Elhoseiny, B. Saleh, A. Elgammal, Write a classifier: zero shot learning using purely textual descriptions, in: ICCV, 2013, pp. 1433–1441.

[10] B. Romera-Paredes, P.H.S. Torr, An embarrassingly simple approach to zero-shot learning, in: ICML, 2015, pp. 2152–2161.
[11] F. Li, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Trans. Pattern Anal. Mach. Intell. 28 (4) (2006) 594–611.
[12] B.M. Lake, R. Salakhutdinov, J.B. Tenenbaum, Human-level concept learning through probabilistic program induction, Science 350 (6266) (2015) 1332–1338.
[13] M. Potthast, B. Stein, M. Anderka, A Wikipedia-based multilingual Retrieval model, in: ECIR, 2008, pp. 522–530.
[14] P. Sorg, P. Cimiano, Exploiting Wikipedia for cross-lingual and multilingual information retrieval, Data Knowl. Eng. 74 (2012) 26–45.
[15] S.L. Smith, D.H.P. Turban, S. Hamblin, N.Y. Hammerla, Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax, 2017.
[16] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, H. Jégou, Word Translation Without Parallel Data, 2018.
[17] Mausam, S. Soderland, O. Etzioni, D.S. Weld, K. Reiter, M. Skinner, M. Sammer, J. Bilmes, Panlingual lexical translation via probabilistic inference, Artif. Intell. 174 (9–10) (2010) 619–637.
[18] M. Paul, A.M. Finch, E. Sumita, How to choose the best pivot language for automatic translation of low-resource languages, ACM Trans. Asian Lang. Inf. Process. 12 (4) (2013) 14.
[19] R. Herbrich, T. Graepel, K. Obermayer, Large Margin Rank Boundaries for Ordinal Regression, MIT Press, Cambridge, MA, 2000.
[20] O. Chapelle, S.S. Keerthi, Efficient algorithms for ranking with SVMs, Inf. Retr. 13 (3) (2010) 201–215.
[21] Y. Song, S. Upadhyay, H. Peng, D. Roth, Cross-lingual dataless classification for many languages, in: IJCAI, 2016, pp. 2901–2907.
[22] Y. Song, S. Mayhew, D. Roth, Cross-lingual dataless classification for languages with small Wikipedia presence, preprint, arXiv:1611.04122.
[23] M.-W. Chang, L. Ratinov, D. Roth, V. Srikumar, Importance of semantic representation: dataless classification, in: AAAI, 2008, pp. 830–835.
[24] Y. Song, D. Roth, On dataless hierarchical text classification, in: AAAI, 2014, pp. 1579–1585.
[25] P.F. Brown, V.J.D. Pietra, P.V. de Souza, J.C. Lai, R.L. Mercer, Class-based n-gram models of natural language, Comput. Linguist. 18 (4) (1992) 467–479.
[26] P. Liang, Semi-Supervised Learning for Natural Language, Master's thesis, Massachusetts Institute of Technology, 2005.
[27] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P.P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.
[28] J. Turian, L.-A. Ratinov, Y. Bengio, Word representations: a simple and general method for semi-supervised learning, in: ACL, 2010, pp. 384–394.
[29] T. Mikolov, W.-T. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: HLT-NAACL, 2013, pp. 746–751.
[30] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, 2013, pp. 3111–3119.
[31] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[32] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, J. Artif. Intell. Res. 34 (1) (2009) 443–498.
[33] Y. Song, D. Roth, Unsupervised sparse vector densification for short text similarity, in: NAACL-HLT, 2015, pp. 1275–1280.
[34] M. Shirakawa, T. Hara, S. Nishio, MLJ: language-independent real-time search of tweets reported by media outlets and journalists, Proc. VLDB Endow. 7 (13) (2014) 1605–1608.
[35] R. Al-Rfou', B. Perozzi, S. Skiena, Polyglot: distributed word representations for multilingual NLP, in: CoNLL, 2013, pp. 183–192.
[36] M. Xiao, Y. Guo, Semi-supervised representation learning for cross-lingual text classification, in: EMNLP, 2013, pp. 1465–1475.
[37] K.M. Hermann, P. Blunsom, Multilingual models for compositional distributed semantics, in: ACL, 2014, pp. 58–68.
[38] M. Faruqui, C. Dyer, Improving vector space word representations using multilingual correlation, in: EACL, 2014, pp. 462–471.
[39] A. Lu, W. Wang, M. Bansal, K. Gimpel, K. Livescu, Deep multilingual correlation for improved word embeddings, in: NAACL-HLT, 2015, pp. 250–256.
[40] S. Upadhyay, M. Faruqui, C. Dyer, D. Roth, Cross-lingual models of word embeddings: an empirical comparison, in: ACL, 2016.
[41] D. Zhang, Q. Mei, C. Zhai, Cross-lingual latent topic extraction, in: ACL, 2010, pp. 1128–1137.
[42] K. Lang, Newsweeder: learning to filter netnews, in: ICML, 1995, pp. 331–339.
[43] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, TACL 5 (2017) 135–146.
[44] A. Raganato, C.D. Bovi, R. Navigli, Automatic construction and evaluation of a large semantically enriched Wikipedia, in: IJCAI, 2016, pp. 2894–2900.
[45] R. Navigli, S.P. Ponzetto, BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artif. Intell. 193 (2012) 217–250.
[46] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, 1998.
[47] J. Camacho-Collados, M.T. Pilehvar, R. Navigli, Nasari: integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities, Artif. Intell. 240 (2016) 36–64.
[48] I. Iacobacci, M.T. Pilehvar, R. Navigli, SensEmbed: learning sense embeddings for word and relational similarity, in: ACL, 2015, pp. 95–105.
[49] C.D. Bovi, A. Raganato, Sew-Embed at SemEval-2017 task 2: language-independent concept representations from a semantically enriched Wikipedia, in: SemEval@ACL, Association for Computational Linguistics, 2017, pp. 261–266.
[50] I. Iacobacci, M.T. Pilehvar, R. Navigli, Embeddings for word sense disambiguation: an evaluation study, in: ACL, 2016.
[51] M.T. Pilehvar, J. Camacho-Collados, R. Navigli, N. Collier, Towards a seamless integration of word senses into downstream NLP applications, in: ACL, 2017, pp. 1857–1869.
[52] D.D. Lewis, Y. Yang, T.G. Rose, F. Li, RCV1: a new benchmark collection for text categorization research, J. Mach. Learn. Res. 5 (2004) 361–397.
[53] Y. Yang, An evaluation of statistical approaches to text categorization, Inf. Retr. 1 (1–2) (1999) 69–90.
[54] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. (ISSN 1532-4435) 9 (2008) 1871–1874.
[55] P. Koehn, Europarl: a parallel corpus for statistical machine translation, in: Machine Translation Summit, 2005, pp. 79–86.
[56] D.C. Howell, Statistical Methods for Psychology, Cengage Learning, 2011.
[57] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[58] P. Prettenhofer, B. Stein, Cross-language text classification using structural correspondence learning, in: ACL, 2010, pp. 1118–1127.
[59] L. Shi, R. Mihalcea, M. Tian, Cross language text classification by model translation and semi-supervised learning, in: EMNLP, 2010, pp. 1057–1067.
[60] G.S. Mann, D. Yarowsky, Multipath translation lexicon induction via bridge languages, in: NAACL, 2001.
[61] T. Cohn, M. Lapata, Machine translation by triangulation: making effective use of multi-parallel corpora, in: ACL, 2007.
[62] M. Utiyama, H. Isahara, A comparison of pivot methods for phrase-based statistical machine translation, in: NAACL-HLT, 2007, pp. 484–491.
[63] H. Wu, H. Wang, Revisiting pivot language approach for machine translation, in: ACL/IJCNLP, 2009, pp. 154–162.
[64] G. Leusch, A. Max, J.M. Crego, H. Ney, Multi-pivot translation by system combination, in: 2010 International Workshop on Spoken Language Translation, IWSLT 2010, Paris, France, December 2–3, 2010, December 2010, pp. 299–306.
[65] M. Yazdani, J. Henderson, A model of zero-shot learning of spoken language understanding, in: EMNLP, 2015, pp. 244–249.
[66] A. Lazaridou, G. Dinu, M. Baroni, Hubness and pollution: delving into cross-space mapping for zero-shot learning, in: ACL, 2015, pp. 270–280.