Low-resource Languages: A Review of Past Work and Future Challenges

Alexandre Magueresse, Vincent Carles, Evan Heetderks
Department of Computer Science and Technology
Tsinghua University, Beijing, China
{maihr19, wenst19, heetderksej11}@mails.tsinghua.edu.cn

Abstract

A current problem in NLP is massaging and processing low-resource languages which lack useful training attributes such as supervised data, number of native speakers or experts, etc. This review paper concisely summarizes previous groundbreaking achievements made towards resolving this problem, and analyzes potential improvements in the context of the overall future research direction.

1 Introduction

In the 1990s, the tools of Natural Language Processing (NLP) experienced a major shift, transitioning from rule-based approaches to statistical-based techniques. Most of today's NLP research focuses on 20 of the 7000 languages of the world, leaving the vast majority of languages understudied. These languages are often referred to as low-resource languages (LRLs), ill-defined as this qualifier can be.

LRLs can be understood as *less studied*, *resource scarce*, *less computerized*, *less privileged*, *less commonly taught*, or *low density*, among other denominations (Singh, 2008; Cieri et al., 2016; Tsvetkov, 2017). In this paper, the term LRL will refer to languages for which statistical methods cannot be directly applied because of data scarcity.

There are many reasons to care about LRLs. Africa and India are the hosts of around 2000 LRLs, and are home to more than 2.5 billion inhabitants. Developing technologies for these languages opens considerable economic perspectives. Also, supporting a language with NLP tools can prevent its extinction and foster its expansion, open the knowledge contained in original works to everyone, or even expand prevention in the context of emergency response. (Tsvetkov, 2017)

This paper provides an overview of the recent methods that have been applied to LRLs as well as underlines promising future direction and points out unsolved questions.

2 Related work

To the best of our knowledge, previous work that conducted reviews on LRLs either aimed attention at specific tasks such as parts-of-speech tagging (Christodoulopoulos et al., 2010), text classification (Cruz and Cheng, 2020) and machine translation (Lakew et al., 2020), or, as the illustrious work on textual analysis (including named entities, parts-of-speech, morphological analysis) (Yarowsky et al., 2001) can tell, trace back to decades ago.

The promotion of LRLs has also been at the core of several conferences and workshops such as LREC [1], AMTA [2] or LoResMT [3].

3 The projection technique

We start by presenting the projection technique, that is central to most NLP tasks applied to LRLs. The rest of our paper examines resource collection with a focus on automatic alignment (section 4), linguistic tasks, which directly take on the projection technique (section 5), speech recognition (section 6), multilingual embeddings (section 7) and their application to machine translation (section 8) and classification (section 9). We conclude by discussing the need for a unified framework to assess the linguistic closeness of two languages, as well as evaluate the ability of a solution to generalize over LRLs, and call further studies to work on more diversified languages.

The projection, or alignment technique has been massively used by the literature since its formalization by (Yarowsky et al., 2001). Indeed, it enables to make the most of the existing annotations of a high-resource language (HRL) and apply it on a LRL, for which annotation is either hard to collect or simply impossible for lack of language

[1] www.lrec-conf.org/
[2] www.amtaweb.org/
[3] https://www.mtsummit2019.com/

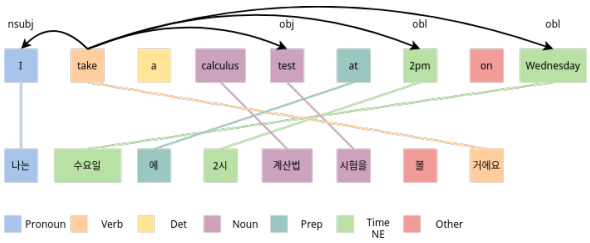

Figure 1: Sample projection from English to Korean

expert.

There are three levels of alignment, document, sentence and word, and each brings about increasing information (in that order) at the cost of more complex annotation. To list a few examples, document-level alignment is useful in information retrieval, sentence-level alignment is at the core of machine translation (section 8), and word-level alignment grounds most linguistic tasks (section 5). Figure 1 provides a sample word-level alignment between English and Korean.

It is difficult to collect corpora aligned at the word level. This is why growing efforts are pursued at automatically aligning corpora, with a focus on word alignment. We further discuss such techniques in section 4.2.

Typically, the ideal one-to-one mapping setting is the exception rather than the rule: there are many cases where the source and target languages exhibit mismatching structures, as can be seen on Figure 1. There are indeed two major prerequisites to ensure annotations can be efficiently projected: i) the two involved languages should be structurally (grammatically) close to each other depending on the task (for example both languages should share the same inventory of tag sets for part-of-speech tagging) ii) the underlying annotations of both languages have to be consistent with each other.

4 Resource collection

By definition, LRLs lack resources that are still necessary for any NLP task. There are two general trends to collecting information for LRLs: i) creating new datasets by annotating raw text ii) gathering raw text and aligning it with a higher-resource language

4.1 Dataset creation

A constant line of efforts has addressed the question of dataset creation, for languages from all around the globe.

First, the REFLEX-LCTL[4] (Simpson et al., 2008) and LORELEI[5](Strassel and Tracey, 2016) projects, conducted by the Linguistic Data Consortium (LDC), released annotated corpora for 13 and 34 languages respectively. LORELEI went a step further by integrating each dataset in a unified framework that includes entity, semantic, part-of-speech and noun phrase annotation.

The challenges faced by both LDC projects served as a lesson for many other programs. (Godard et al., 2017) produced a speech corpus that can be used for speech recognition and language documentation. (Marivate et al., 2020a,b) curated a collection of news headlines for two South African languages, as well as established baselines for the classification task. This work showed what was possible to achieve using very limited amounts of data and should be an example to guide future work in collecting new dataset for LRLs.

Future work The aforementioned studies proved greatly innovative in the sources they examined to built their datasets. Social media, mobile applications but also governmental documents can all be taken advantage of to generate textual content.

4.2 Automatic alignment

In the typical scenario, automatic alignment requires language expertise and is time-consuming. However, since most techniques involving LRLs rely on aligned corpora, it has been the research focus of many recent studies.

Word-level alignment The most elementary alignment scope is at the word-level ; this task is most commonly referred to as lexicon induction. The inverse consultation (IC) method (Saralegi et al., 2011), that builds and compares both dictionaries $A \rightarrow C$ and $C \rightarrow A$ from dictionaries $A \leftrightarrow B$ and $B \leftrightarrow C$, has been used in varied contexts, but fails to model lexical variants and polysemy (Figure 2, “joie” will be translated as “felio” whereas it should not). The distributional similarity (DS) method examines the context of a word to further improve the IC algorithm.

The more recent cognate equivalence assumption, that claims that two words sharing similar writing, meaning and etymology share all their meanings, significantly helped develop lexicon

[4]Research on English and Foreign Language Exploitation – Less Commonly Taught Languages

[5]LOw-REsource Languages for Emergent Incidents

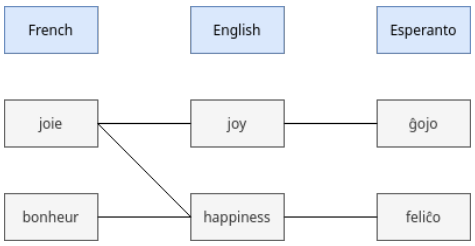

Figure 2: Ambiguity problem of the pivot technique

and synonym dictionaries (Nasution et al., 2016, 2017). Given two dictionaries $A \rightarrow B$ and $B \rightarrow C$, other work introduced constraints on the similarity of languages A and C and modeled the structures of the input dictionaries as a Boolean optimization problem, in order to build dictionary $A \rightarrow C$ (Wushouer et al., 2015).

Sentence-level alignment Particularly useful for machine translation, sentence-level alignment has recently gained momentum. The general method consists in two steps: devising a similarity score between two sentences and aligning sentences based on these scores. A great variety of methods aimed to evaluate the distance between two sentences. Since the release of the WMT16 dataset [6], translating the source sentence into the target language and matching n-grams (Dara and Lin, 2016; Gomes and Lopes, 2016; Shchukin et al., 2016), or using cosine similarity on tf-idf vectors (Buck and Koehn, 2016; Jakubina and Langlais, 2016) has become popular. Even if word-to-word translations can be applied to LRLs, these systems often require higher-quality translation. Consequently, another trend preferably relies on multilingual embedding such as (Artetxe and Schwenk, 2019) and derive a similarity score with the word mover distance (WMD) (Huang et al., 2016; Mueller and Lal, 2019).

Document-level alignment Aligning corpora at the document-level is useful for text classification, translation or multilingual representations. Based on sentence embedding and more relaxed formulations of the Earth mover's distance, (El-Kishky and Guzmán, 2020) obtained state-of-the-art alignment on low- and mid-resource languages.

Future work More emphasis should be put on sentence distance evaluation, since it directly influences current results for document alignment. Other solutions to improve document alignment could include designing new weighting schemes to account for significant differences in document sizes. Finally, heuristics could help cut down the complexity of document alignment methods, such as relative position in the document or sentence length.

5 Linguistic tasks

The "linguistic" tasks we consider in this paper are mostly related to grammar modeling, and their applications serve the interest of linguistics and research rather than any commercial purpose (such as machine translation, speech recognition, summarization and many other).

5.1 Part-of-speech tagging

Solving the part of speech (POS) tagging task in an unsupervised fashion essentially amounts to a clustering task. Once every word is assigned to one or several clusters, each cluster needs to be grounded, or mapped to its corresponding POS. Grounding fatally requires a minimum annotation.

There are many clustering algorithms, of which Brown (Brown et al., 1992) is one of the earliest formulation. It has yet been shown that this simple clustering technique is often the one that leads to best performance (Christodoulopoulos et al., 2010). Another very common technique inherits the hidden Markov model (HMM) and views the POS tagging task as a sequence-to-sequence problem. By training on a "parent" language for which annotations are available, it is even possible to operate the grounding step without annotation for the LRL, provided that the tags are standardized among both languages (Buys and Botha, 2016; Cardenas et al., 2019). Both works showed that the choice of the "parent" language is difficult, because typologically close languages do not always work best in practise, even when relying on expert features provided by the WALS [7] features. Instead, it advocates combining several parent languages to better leverage word-order patterns, and shirk the question of parent language selection.

Supervised (or semi-supervised) methods are also applied to the POS tagging task, using projection to cope with the lack of annotations. Earliest methods restricted their application to closely-related languages by directly tagging the target sentence the same way as the source sentence (also

[6] http://www.statmt.org/wmt16/

[7] www.wals.info

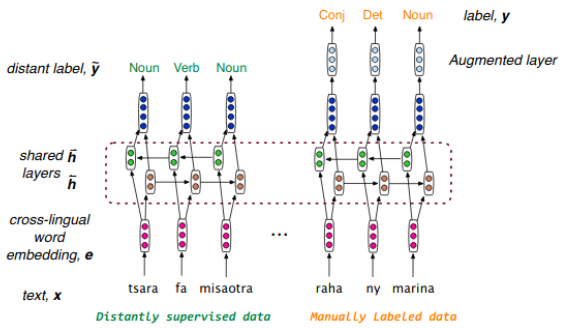

Figure 3: One model architecture for POS-tagging, from (Fang and Cohn, 2017)

including class rebalancing to account for minor linguistic differences) (Yarowsky et al., 2001). Long short-term memory networks (LSTM) based on multilingual embeddings proved efficient when available annotations were in larger quantities, proved (Zennaki et al., 2015).

POS-tagging as a classification problem was also attacked with hand-crafted features on very limited amounts of annotations, backed with reasonable help of projection on the English language (Duong et al., 2014). The proposed model is a softmax classifier, first trained on projected annotations, and then adjusted on ground-truth tags. Two other formulations relying on bidirectional LSTMs further improved tagging accuracy on some languages (Fang and Cohn, 2016, 2017). As can be seen on Figure 3, the output of the BiLSTM model is either directly used for the distantly annotated data, or first passed through a projection layer to be evaluated on the ground truth tags of the LRL. The two formulations differ in the way the hidden states of the BiLSTM are projected (either matrix multiplication (Fang and Cohn, 2016) or MLP (Fang and Cohn, 2017)).

Future work Most projected annotations currently stem from English corpora. One direction for future work would be relying one or several other HRLs to transfer annotations from. Also, it might be interesting to find a way to combine the information of multiple LRLs simultaneously, especially if they are closely-related. Besides, a more robust evaluation metric has to be adopted by unsupervised methods for POS tagging. (Christodoulopoulos et al., 2010) suggests using the V-measure, analogous to the F-score metrics.

5.2 Dependency parsing

This section first summarizes two important findings that greatly influenced the following studies. Then, it shows how parameter sharing and zero-shot transfer are profitable to dependency parsing. Both techniques are facilitated by unified annotation schemes, such as the Universal Dependencies (UD) dataset [8].

Two major findings Dependency parsing has traditionally built on POS-tagging, but in the context of LRLs, gold POS tags are not available. Replacing POS information by inferred word clusters is often more beneficial than trying to predict gold POS (Spitkovsky et al., 2011). Allowing words to belong to different clusters in different contexts indeed helps model polysemy. Another study showed that transferring from multiple sources is generally preferable to single source when languages share grammatical similarities (McDonald et al., 2011) ; not because more data is leveraged but because the model is thus exposed to a variety of patterns.

Parameter sharing There are even some cases where distant supervision through jointly learning dependency parsing on a source and target languages leads to more accurate performance for the target language, than relying on supervised data for the target language (Duong et al., 2015a,b). These two papers let their model share parameters across two models, and only the embedding layer is language-dependent.

Zero-shot transfer Transition-based LSTMs predicting sequential manipulations on a sentence (reduce left or right, shift) taking cluster and POS information as well as embedding as input outperforms previous cross-lingual multi-source model transfer methods (Ammar et al., 2016). Following the multi-source transfer guideline, (Agić et al., 2016) projects annotations weighted by alignment probability scores. This weighted scheme averaged on a large number of languages reached unprecedented unlabeled attachment score (UAS) for more than 20 languages. (Wu et al., 2018) establishes that carefully selecting source languages allows for efficient direct transfer learning from models that perform well on HRLs. This naturally raises the question of how to optimally select proper source languages given a LRL.

[8] www.universaldependencies.org

Future work Several studies suggest learning POS and dependencies in a joint fashion would incite major advances in both tasks. (Lim et al., 2020) recently presented a multi-view learning framework that jointly learns POS tagging and parsing. Both token and character embeddings feed two independent LSTMs, the hidden state of which are used as input as another central LSTM. The outputs of the three models are then combined and co-trained on both tasks.

Besides, sharing parameters across closely-related languages could give rise to further studies on larger scales and more diverse language sets. More generally, the question of language "closeness" is a topical issue that needs careful consideration.

5.3 Named entity recognition, typing and linking

Extracting (recognizing, NER) named entities from a text, classifying them (typing, NET) into categories (such as location, name or organization) and building bridges across them (linking, NEL) is of capital importance in information retrieval, recommendation system and classification. In the context of LRLs, these three tasks have been the subject of increasing research.

NER and NET Both tasks are often jointly learnt as a classification task which includes a category for wrods that are not named entities. Projection has been widely used in this task, either to learn a model on a HRL and applying it to a LRL (Zamin, 2020), or to train a classifier on the projected annotations from a HRL (Yarowsky et al., 2001). The early solutions to solve this task mainly involved Hidden Markovian Models (HMM), but they fail to handle named entity phrases and cannot model non-local dependencies along the sentence (Yarowsky et al., 2001). A transition towards Conditional Random Fields (CRF) helped answer these two issues, at the cost of hand-made and language-dependent features (Saha et al., 2008; Demir and Özgür, 2014; Littell et al., 2016). The features not only involved prefixes and suffixes, but also more language-specific morphology induction methods and transcriptions in the international phonetic alphabet.

More recent work replaced costly feature engineering with embeddings, that are learned from BiLSTMs (Cotterell and Duh, 2017; Suriyachay et al., 2019). A CRF is placed at the end of the pipeline to handle the classification. A very different approach, hypothesizing that NE of the same type are embedded in the same region, showed that viewing the typing task as a clustering approach based on embeddings can outperform CRF architectures (Das et al., 2017). Phrases can easily be taken into account by this method. Finally (Mbouopda and Yonta, 2020) takes advantage of the low frequency of NEs in any document to derive a novel embedding for corpora aligned at the sentence-level. A traditional neural classifier is then fed with these embedding to perform classification. This contribution has the advantage of only requiring sentence alignment, and can be applied to other classification tasks where a few classes are more prevalent than many other.

NEL Cross-lingual NEL (XEL) generally consists in two steps: candidate generation and candidate ranking. External sources of knowledge such as Wikipedia or Google Maps often help link or disambiguate entities (Gad-Elrab et al., 2015). A feature-based neural model further improved by designing a new feature combination technique and (Zhou et al., 2019) proposed a new scheme to integrate two candidate generation methods (look-up- and neural-based), suggest a set of language-agnostic features and devise a non-linear combination method. This resulted in significant improvement in end-to-end NEL. (Zhou et al., 2020) replace the LSTM model by an n-gram bilingual model to solve the sub-optimal string modeling. Finally, (Fu et al., 2020) relies on log queries, morphological normalization, transliteration and projection as a comprehensive improvement over even supervised methods.

Future work As (Zamin, 2020) reported, performance of NER systems is highly currently domain-specific ; some papers incentivize designing cross-domain techniques. The literature mentioned that conceiving a new metrics to compare cross-lingual embedding is promising avenue to XEL. Transferring from languages other than English would further allow to check the robustness of the currently developed methods.

5.4 Morphology induction

Morphology induction aims to align inflected word forms with their root form. This task unifies lemmatization and morphological paradigm completion, which is the reverse process, that is finding all inflected forms from a root word.

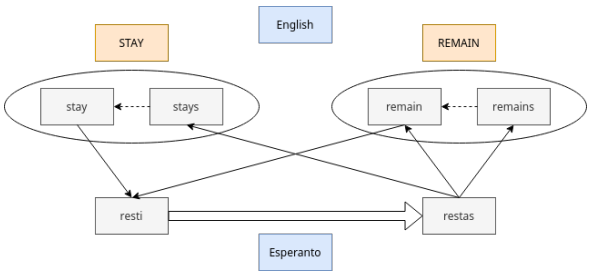

Figure 4: Morphology induction for Esperanto through English

The earliest method relied on word alignment between a source HRL and a target LRL and an existing morphology induction system in the source language (Yarowsky et al., 2001). Essentially, a word in the target language is mapped to its counterpart in the source language *via* word-alignment, which redirects to its root form thanks to the morphology system. Word-alignment enables to map the source root form back into the target language to produce the desired output (Figure 4).

A more advanced method makes the assumption that the word inflection occurs at the end of the word to derive a trie-based approach. If the inflected i and root r word share a longest common substring s such that $i = s.\tilde{i}$ and $r = s.\tilde{r}$, then the probability that i comes from r is defined by

$$P(i|r) = \sum_{k=0}^{|r|} \lambda_k P(\tilde{r} \rightarrow \tilde{i}|r_{<k})$$

For example, $P(replies|reply) = \lambda_0 P(y \rightarrow ies) + \lambda_1 P(y \rightarrow ies|y) + \lambda_2 P(y \rightarrow ies|ly) + \lambda_3 P(y \rightarrow ies|ply) + \ldots$. Other methods to lemmatization involved learning edit-based strategies (copy, delete, replace) through LSTMs models like (Makarov and Clematide, 2018), the sequential nature of which strikingly outperforms character aligners.

Another more recent line of research has aimed to create morphological paradigms, that is the set of all rules that can be applied on a root form given its part-of-speech. The unsupervised process consists in three steps: candidate search, paradigm merging and generation. While a reductive approach relies on tries (Monson et al., 2007), a more robust method involves edit trees (Jin et al., 2020) to model the transition between two words. Surrounding context is used to merge paradigms, and a transducer chooses from the different available shifts for a given slot to create the final output. For example in English, the third person of the present tense can be obtained by adding "s" or "es". The transducer chooses to apply "es" to the root "miss" rather than "s".

Future work This last work achieves strong performance in a very low-resource setting under the hypothesis that a morphological shift only appears in one paradigm. This is not always true (for example in English adding an "s" both enables to transition from singular to plural for nouns, and to create the third person verbal from from the infinitive). Future directions could include releasing this limiting assumption as well as making use of word embedding to merge paradigms and exploring other transduction models.

6 Speech recognition

The goal of large vocabulary continuous speech recognition (LVCSR) is to derive the meaning of a verbal message by identifying embedded phones. Applying LVSCR on LRLs primarily concerns how to exploit various language components such as phones and syllables while also employing transfer learning from high-resource secondary languages to a target LRL.

Multilayer perceptrons A traditional approach for multilayer perceptron (MLP) networks in speech recognition is incorporating phonal training data from other HRLs and extracting tandem features for LRL classification. However, this originally required that data from all languages be transcribed using a common phoneset. The work in (Thomas et al., 2012) instead presented an MLP model that completed the same task without mapping each language to the same phoneset. As shown in Figure 5, this is accomplished first by training a 4-layer MLP with layers d, h_1, h_2, and p on HRLs. The last layer p is then removed and replaced with a single-layer perceptron q whose weights have been pre-trained using only limited data from the target LRL. The entire 4-layer model was lastly trained again using only target LRL data. This approach is roughly 30 percent more accurate than the baseline common-phoneset model.

Hidden Markov models Another popular speech recognition method is the hidden Markov model (HMM), which is a subclass of unsupervised dynamic Bayesian networks. An HMM in speech recognition essentially works by learning

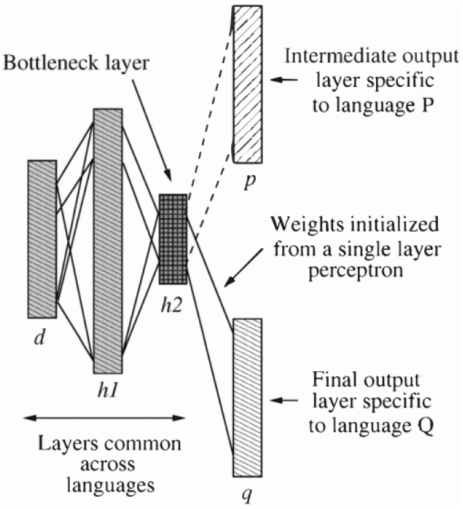

Figure 5: MLP model proposed by (Thomas et al., 2012) that replaces the last layer of a network trained on HRL 'P' with a single layer pre-trained on LRL 'Q'

the probabilities for next sequential sounds given a particular sound in training speech. It is then made to confirm sequential sounds in speech on a phone, syllable, and word level. From an low-resource standpoint, HMM topology and the scope of speech modelling units are major points of consideration for ensuring success. For example, (Fantaye et al., 2019) adopted a particular LRL's relatively few possible consonant-vowel syllables as modelling units and iterated through different HMM architectures. A deep-neural-network-based HMM using transfer learning from a language similar to the said LRL achieved greatest results. The same transfer learning paradigm for HMM-based speech recognition in LRLs also explored in (Adams et al., 2017), which uses English as a source language for Turkish.

Future work One literature focus is performance of name and place entity detection in speech. One avenue of future work must address how to improve the detection accuracy of these speech components. Another concern is the lack of homogeneity among speech training data, which have various noise, speakers, accents, emotions, etc. that are difficult to standardize and create lots of variance. Further research should yield methods to filter and normalize training data.

7 Embeddings

Embeddings are the key to accurate and efficient NLP models. They represent a model's understanding of sentence structure and word meaning. However, training embeddings is a resource-heavy process and is not compatible with the data scarcity of LRLs. Hence, to avoid a random initialization of embeddings, different approaches are being used: multilingual representations, transfer learning for structurally similar languages, or even data augmentation.

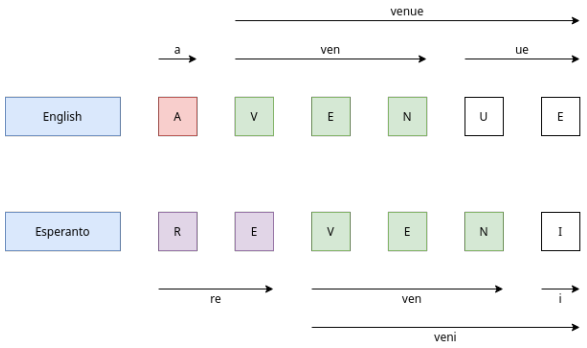

Figure 6: The radical "ven" is quite common in Esperanto, thus it could be integrated in the English-Esperanto segmentation, and benefit a few English word including "venue", "avenue", "advent" and others (Nguyen and Chiang, 2017)

7.1 Data Augmentation

Byte Pair Encoding segmentation, known as BPE segmentation, extracts additional data from the language corpus in order to generate the embeddings (Nguyen and Chiang, 2017). The motivation behind BPE segmentation to take advantage of shared substrings: words from both dictionaries are broken down into subwords and then the most common subwords are kept as roots. (Figure 6). This means that each word is therefore seen as a sequence of tokens (or subwords) which increases the amount of overlap between both vocabularies. This technique is particularly useful for LRLs as it improves the transferability of the embeddings. Indeed, having more tokens in common between the parent HRL and the target LRL means there are more overlapping words to align the embeddings during transfer.

To combat the training data scarcity, another approach is to transform a LRL into a HRL using data augmentation. A technique used in computer vision was adapted to NLP by (Fadaee et al., 2017) (Figure 7).

The idea is to create additional sentences by replacing a single word in the sentences of the dataset. Still, the words replaced for data augmentation must fit the sentence structure properly (e.g.

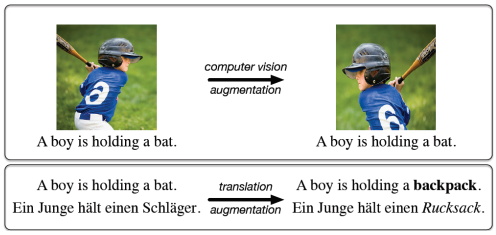

Figure 7: Comparison between Computer Vision and Low-Resource Languages data augmentation (Fadaee et al., 2017).

verb replacing another verb) in order for the generated sentence to be usable during training. However, even if the sentence structure is preserved, its meaning may be altered. This is problematic for multiple LRL domains and is rarely used outside of machine translation models. Indeed, for NMT, models are built to translate even meaningless sentences (e.g.: Boats are delicious!) as they only require proper word pairs translations and correct sentence structure.

Future work Although Data Augmentation techniques proved successful to combat several LRLs, current approaches struggle for extremely low-resource languages. In fact, to properly handle data augmentation, precise knowledge of sentence structure, grammar and words translation pairs is required in order to assert the correctness of the generated sentence.

7.2 Multilingual Embeddings

Multilingual embedding-based models have been introduced for zero-shot cross-lingual transfer. For instance, using a single Bidirectional Long Short Term Memory (BiLSTM) as the model encoder (Artetxe and Schwenk, 2019) allows to learn a classifier after training the multilingual embeddings. Therefore, the embeddings can be transferred to any of the languages the NMT is designed for without any structural modifications.

Another multilingual embedding-based model architecture uses multilingual BERT to perform accurate zero-shot cross-lingual model transfer (Pires et al., 2019). M-BERT proved effective at transferring knowledge across languages with different scripts (no lexical overlap), which proves that the model does capture multilingual representations into embeddings. These can later serve as a basis for multilingual tasks.

8 Machine translation

A Neural Machine Translation model, known as NMT model, aims to translate a sentence from a source language to a target language. NMT models are built using two separate but connected models: an encoder and a decoder. The encoder's goal is to break down the sentence logic and word pairs, and then store this information as embeddings. Afterwards, the decoder generates the translated sentence based on those embeddings.

This technology has proved to be highly accurate and reliable for translating between two HRLs. Yet, NMT models require large corpora of training data, based on translations or annotated data crafted by language experts. For common languages such as English, Spanish or French, the data used for training has already been processed in large quantities. However, when it comes to less common languages and dialects, there is little to none translated data available. That is why specific architectures and methods are required to handle LRL NMT. The state-of-the-art techniques and models are introduced in the following sections.

8.1 Transfer Learning

Transfer Learning (Zoph et al., 2016) relies on the fact that NMT achieves great results for HRLs. Indeed, it uses a two-model architecture: a parent model which is a standard NMT model trained on a pair of HRLs (e.g. French-English or English-Spanish) and a child model which is the desired translation model between the source and target languages. More specifically, the parent model is trained using a standard corpus and then, the embeddings of the parent model are used to initialize those of the child model. This way, the embeddings of the child model are not randomly initialized and it can be trained more efficiently even with a small amount of data.

Future work Transfer Learning works best when the parent and child languages have similar lexicon and grammar. Strategies to combat this issue still have to be developed, such as using an ordering model (Murthy et al., 2018). The goal of such a model is to reorder the sentences of the source language to match the structure of the target language. Having the same sentence structure allows for a more accurate transfer of embeddings

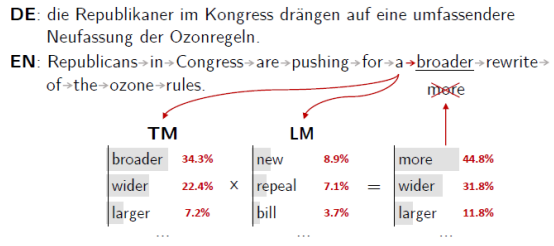

Figure 8: Example of an incorrect prediction when the LM and NMT model disagree (Baziotis et al., 2020).

as similar words will have similar positions in sentences.

8.2 Prior Models

A second approach to cope with LRLs consists in introducing a prior model during training. Indeed, the Zero-Shot Translation technique (Kumar et al., 2019) allows to train the encoder model using multiple languages or dialects simultaneously. This model is able to capitalize on the already learned languages pairs to translate between unseen language pairs. For instance, if the model has been trained for English - French and French - Chinese pairs then it can already handle English - Chinese translation. Thus avoiding the rebuilding of the NMT system for every new language pair.

Further works have replaced the Zero-Shot Translation model by a Language Models (LM) (Baziotis et al., 2020). The LM adds a regularization term that pushes likelier output distributions to the NMT. This means that, out of the top predictions output by the NMT, only the top word amongst the ones which are validated by the LM will be selected. Hence, this can be seen as a knowledge distillation technique where the LM is teaching the NMT about the target language.

Future work Language Model-based NMT, however, is still exposed to incorrect predictions in some cases where the LM and the NMT model disagree on a prediction, even when the NMT model predicted correctly (Figure 8).

8.3 Multilingual Learning

Multilingual Learning extends the transfer learning techniques in multilingual environments. The idea is to build an NMT model using a universal shared lexicon and a shared sentence-level representation, which is trained using multiple LRLs. One of the first approaches (Gu et al., 2018a) consists in using two additional components compared to the traditional transfer learning techniques. The first one is a universal lexical representation to design the underlying word representation for the shared embeddings. The second is a mixture of language experts to deal with the sentence-level sharing. This system is used during the encoding and works as a universal sentence encoder. It allows to transfer the learned embeddings for a given language pair into the universal lexical representation.

A Model-agnostic meta-learning (MAML) algorithm applied to low-resource machine translation allows to view language pairs as separate tasks (Gu et al., 2018b). This technique results in faster and more accurate training of the vocabulary. Although, as MAML was initially designed for deep learning purposes, it cannot handle tasks with mismatched input and output. That is why it is only applied to Universal NMT models.

Language Graph A rather structurally different approach introduces the concept of Language Graph for Multilingual Learning (He et al., 2019). Vertices represent the various LRLs and edges represent the translation word pairs. Moreover, for each edge, a weight score is assigned to denote the accuracy of the translation pair. Then, a distillation algorithm is used to maximize the total weight (accuracy) of the model, using forward and backward knowledge distillation to boost accuracy. The main advantage of such a graph is that there exist multiple translation paths from one word to another. (e.g. English to Spanish and English to French to Spanish). For LRLs, the direct translation path generally has low accuracy due to the lack of available parallel data. However, there are some languages for which the translation pair will have high accuracy and this resulting path will be put to use.

9 Classification

Sentiment analysis Advanced opinion mining models designed for the English language do not work well on LRLs that have vastly different grammar, unstructured format, and little applied NLP research or resources. From a machine learning standpoint, one alternative is adopting lighter, less resource-dependent baseline models that are still successful with the English language and modifying them to maximize success with a particular LRL. For example, (Al-Sallab et al., 2017) adopted a recursive auto encoder (RAE) baseline

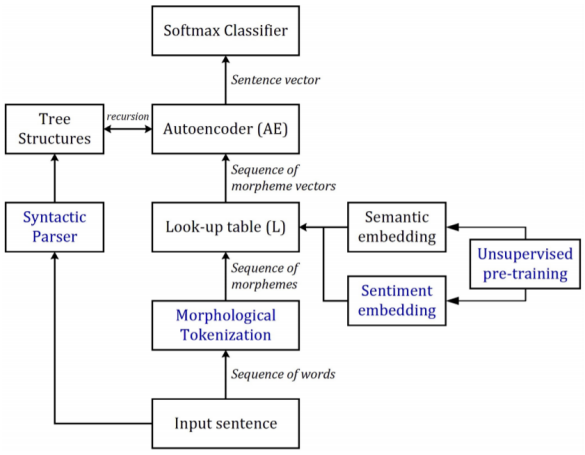

Figure 9: Aroma framework proposed by (Al-Sallab et al., 2017), with augmented portions in blue

model for Arabic and augmented it with morphological tokenization, a sentiment-extracting neural network, an unsupervised pre-training block to improve sentiment embedding initialization, and a phrase structure parser to generate parse trees (Figure 9).

An alternative solution is the lexicon-based approach, which manually classifies words in a sentence based on a list of sentimentally-classed words. The lexicon-based method, various machine learning methods such as K Nearest Neighbor, Naive Bayes, and Decision Tree, as well as hybrid architectures between the two are applied on the English and Urdu languages in (Azam et al., 2020).

Data expansion One way to lift LRLs out of the "low-resource" category is obviously to increase the quality and quantity of available supervised classification data. One such approach is to expand existing data through adversarial distortion of text attributes, or by extracting more features using a transfer learning technique that surmount pre-trained layers from recognition systems for multiple common languages with new trainable layers (Qi et al., 2019). A more direct approach is to manually compile a corpus dataset that includes multi-label text classifications and pre-training language models as a platform for further NLP work for respective LRLs as done in (Cruz and Cheng, 2020) for the Filipino language.

Miscellaneous There are many more LRL classification sub-topics. One such avenue is text readability classification which, for example, can automate quality analysis and pinpoint areas requiring edit in LRL textbooks using lexical, entropy-based, and divergence-based features (Islam et al., 2012).

Pattern recognition is another problem for LRLs due to the lack of NLP study on coping with less common textual attributes. The work in (Abliz et al., 2020) has addressed impediments such as vowel weakening and suffix-based morphological changes for the Uyghur language through a proposed algorithm that performs pattern matching using syllable features.

Future work Although the above literature had successful approaches for addressing the LRL classification problem in their respective areas, the overall research direction is still rooted in the fundamental understudy and lack of experience with LRLs. Future classification work seems to be be focused on two general categories. One line will be further study of LRL morphological traits and grammar patterns to increase model performance. Another branch of research will focus on applying existing classification models to use as a benchmark to improve off of for more novel methods.

10 Discussion

On top of the low-level examination we conducted for various NLP tasks, we would like to summarize two desiderata recurrently noted the literature: collecting new datasets for more diverse languages and devising a closeness index for languages.

Datasets diversity A few papers collected new datasets in innovative ways, that we believe should be further put forward. Extracting news headlines or comments from social media (Marivate et al., 2020a), relying on mobile applications to gather audio extracts and annotations (Godard et al., 2017), as well as relying on governmental sources, constitute new alleys to dataset creation. Finally, we would like to mention the Tatoeba [9] project, that is a collaborative, open and free collection of aligned sentences and translations in more than 350 languages.

Closeness index for languages Throughout our review, we encountered a fair amount of papers that reported the difficulty to select language pairs that allow for smooth transfer or alignment. As quoted by (Nasution et al., 2017), the Automated Similarity Judgment Program (ASJP) (Wichmann

[9] www.tatoeba.org

and , eds.) [10] has collected a word list based on the Swadesh list for more than 9500 languages and dialects. This allows for a morphological and lexical comparison, yet cannot help match grammatical differences across languages. That is why we advocate the design of a task-specific linguistic distance that would model both the morphological and grammatical aspects of a language, and, given a target language, would guide the choice of the optimal language to transfer from.

11 Conclusion

A review of over 60 LRL-related papers has yielded a general idea of this field's most recent work. In an area that concerns a fundamental lack of data, a primary trend is expanding LRLs datasets while also applying augmentation methods and transfer learning techniques from other languages in a manner that copes with their differences. Future work will highly involve improving the quality of LRLs data, taking advantage of linguistic patterns/similarities, designing more robust learning models, and increasing the reliability of evaluation methods.

References

Wayit Abliz, Maihemuti Maimaiti, Hao Wu, Jiamila Wushouer, Kahaerjiang Abiderexiti, Tuergen Yibulayin, and Aishan Wumaier. 2020. Research on uyghur pattern matching based on syllable features. *Information*, 11(5):248.

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–20.

[10] www.asjp.clld.org/

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Nazish Azam, Bilal Tahir, and Muhammad Amir Mehmood. 2020. Sentiment and emotion analysis of text: A survey on approaches and resources. *LANGUAGE & TECHNOLOGY*, page 87.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Christian Buck and Philipp Koehn. 2016. Quick and reliable document alignment via tf/idf-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678.

Jan Buys and Jan A Botha. 2016. Cross-lingual morphological tagging for low-resource languages. *arXiv preprint arXiv:1606.04279*.

Ronald Cardenas, Ying Lin, Heng Ji, and Jonathan May. 2019. A grounded unsupervised universal part-of-speech tagger for low-resource languages. *arXiv preprint arXiv:1904.05426*.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised pos induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584. Association for Computational Linguistics.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96.

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068*.

Aswarth Abhilash Dara and Yiu-Chang Lin. 2016. Yoda system for wmt16 shared task: Bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 679–684.

Arjun Das, Debasis Ganguly, and Utpal Garain. 2017. Named entity recognition with word embeddings and wikipedia categories for a low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(3):1–19.

Hakan Demir and Arzucan Özgür. 2014. Improving named entity recognition for morphologically rich languages using word embeddings. In *2014 13th International Conference on Machine Learning and Applications*, pages 117–122. IEEE.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015a. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015b. A neural network model for low-resource universal dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? a case study of multilingual pos tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897.

Ahmed El-Kishky and Francisco Guzmán. 2020. Massively multilingual document alignment with cross-lingual sentence-mover's distance. *arXiv preprint arXiv:2002.00761*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource pos tagging using cross-lingual projection. *arXiv preprint arXiv:1607.01133*.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. *arXiv preprint arXiv:1705.00424*.

Tessfu Geteye Fantaye, Junqing Yu, and Tulu Tilahun Hailu. 2019. Syllable-based speech recognition for a very low-resource language, chaha. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 415–420.

Xingyu Fu, Weijia Shi, Zian Zhao, Xiaodong Yu, and Dan Roth. 2020. Design challenges for low-resource cross-lingual entity linking. *arXiv preprint arXiv:2005.00692*.

Mohamed H Gad-Elrab, Mohamed Amir Yosef, and Gerhard Weikum. 2015. Named entity disambiguation for resource-poor languages. In *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 29–34.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noël Kouarata, Lori Lamel, Hélene Maynard, Markus Müller, et al. 2017. A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*.

Luís Gomes and Gabriel Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018a. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. 2018b. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

Tianyu He, Jiale Chen, Xu Tan, and Tao Qin. 2019. Language graph distillation for low-resource machine translation. *arXiv preprint arXiv:1908.06258*.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pages 4862–4870.

Zahurul Islam, Alexander Mehler, and Rashedur Rahman. 2012. Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

Laurent Jakubina and Philippe Langlais. 2016. Bad luc@ wmt 2016: a bilingual document alignment platform based on lucene. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 703–709.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya D McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. *arXiv preprint arXiv:2005.00970*.

Rashi Kumar, Piyush Jha, and Vineet Sahula. 2019. An augmented translation technique for low resource language pair: Sanskrit to hindi translation. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 377–383.

Surafel M Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages. *arXiv preprint arXiv:2003.14402*.

KyungTae Lim, Jay Yoon Lee, Jaime Carbonell, and Thierry Poibeau. 2020. Semi-supervised learning on meta structure: Multi-task tagging and parsing in low-resource scenarios. In *Proceedings of the AAAI Conference*.

Patrick Littell, Kartik Goyal, David R Mortensen, Alexa N Little, Chris Dyer, and Lori Levin. 2016. Named entity recognition for linguistic rapid response in low-resource languages: Sorani kurdish and tajik. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 998–1006.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020a. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2003.04986*.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020b. Low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2004.13842*.

Michael Franklin Mbouopda and Paulin Melatagia Yonta. 2020. Named entity recognition in low-resource languages using cross-lingual distributional word representation. *HAL archives ouvertes*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics.

Christian Monson, Jaime Carbonell, Alon Lavie, and Lori Levin. 2007. Paramor: Finding paradigms across morphology. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 900–907. Springer.

Aaron Mueller and Yash Kumar Lal. 2019. Sentence-level adaptation for low-resource neural machine translation. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 39–47.

V Murthy, Anoop Kunchukuttan, Pushpak Bhattacharyya, et al. 2018. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1811.00383*.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2016. Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3291–3298.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2017. A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(2):1–29.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *arXiv preprint arXiv:1708.09803*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Zhaodi Qi, Yong Ma, and Mingliang Gu. 2019. A study on low-resource language identification. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1897–1902. IEEE.

Sujan Kumar Saha, Sanjay Chatterji, Sandipan Dandapat, Sudeshna Sarkar, and Pabitra Mitra. 2008. A hybrid named entity recognition system for south and south east asian languages. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Xabier Saralegi, Iker Manterola, and Inaki San Vicente. 2011. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics.

Vadim Shchukin, Dmitry Khristich, and Irina Galinskaya. 2016. Word clustering approach to bilingual document alignment (wmt 2016 shared task). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 740–744.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources. *Collaboration: interoperability between people in the creation of language resources for less-resourced languages*, 7.

Anil Kumar Singh. 2008. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Valentin I Spitkovsky, Hiyan Alshawi, Angel X Chang, and Daniel Jurafsky. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1281–1290. Association for Computational Linguistics.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

Kitiya Suriyachay, Thatsanee Charoenporn, and Virach Sornlertlamvanich. 2019. Thai named entity tagged corpus annotation scheme and self verification. In *Proceedings of the 9th Language & Technology Conference (LTC2019)*.

Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky. 2012. Multilingual mlp features for low-resource lvcsr systems. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4269–4272. IEEE.

Yulia Tsvetkov. 2017. Opportunities and challenges in working with low-resource languages. Carnegie Mellon University.

Eric W. Holman Wichmann, Sren and Cecil H. Brown (eds.). 2020. The asjp database (version 19).

Yingting Wu, Hai Zhao, and Jia-Jun Tong. 2018. Multilingual universal dependency parsing from raw text with low-resource language enhancement. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 74–80.

Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. 2015. A constraint approach to pivot-based bilingual dictionary induction. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 15(1):1–26.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Norshuhani Zamin. 2020. Projecting named entity tags from a resource rich language to a resource poor language. *Journal of Information and Communication Technology*, 12:121–146.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. 2015. Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle. In *TALN 2015*.

Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019. Towards zero-resource cross-lingual entity linking. *arXiv preprint arXiv:1909.13180*.

Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association for Computational Linguistics*, 8:109–124.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.