Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 271–276
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Building Language Models for Morphological Rich Low-Resource Languages using Data from Related Donor Languages: the Case of Uyghur

Ayimunishagu Abulimiti, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany
{ay.abulimiti, tanja.schultz}@uni-bremen.de

Abstract

Huge amounts of data are needed to build reliable statistical language models. Automatic speech processing tasks in low-resource languages typically suffer from lower performances due to weak or unreliable language models. Furthermore, language modeling for agglutinative languages is very challenging, as the morphological richness results in higher Out Of Vocabulary (OOV) rate. In this work, we show our effort to build word-based as well as morpheme-based language models for Uyghur, a language that combines both challenges, i.e. it is a low-resource and agglutinative language. Fortunately, there exists a closely-related rich-resource language, namely Turkish. Here, we present our work on leveraging Turkish text data to improve Uyghur language models. To maximize the overlap between Uyghur and Turkish words, the Turkish data is pre-processed on the word surface level, which results in 7.76% OOV-rate reduction on the Uyghur development set. To investigate various levels of low-resource conditions, different subsets of Uyghur data are generated. On the smallest subset including only 100 Uyghur utterances, a word-based language model trained with bilingual Uyghur-Turkish data achieved 98.10% relative perplexity reduction over the language models trained with Uyghur data only. Morpheme-based language models trained with bilingual data achieved up to 40.91% relative perplexity reduction over the language models trained only with Uyghur data.

Keywords: Multilingual, Low-resource, Language modeling, Agglutinative languages, GlobalPhone

1. Introduction

A language model is one of the main components of Automatic Speech Recognition (ASR) systems, which significantly impacts the overall recognition performances. To build reliable language models, very large amounts of text data are required. However, even with large corpora, the construction of reliable language models is very challenging for morphological rich languages, since the large vocabulary leads to high Out-Of-Vocabulary (OOV) rates. This problem becomes even more dramatic if only few data resources are available for a language in question. In this paper we address the example of Uyghur, which combines both challenges, i.e. Uyghur has a rich morphology that primarily uses agglutination and Uyghur belongs to the category of low-resource languages.

A common approach in language modeling to overcome high OOV rates in agglutinating languages is the use of subwords or morphemes as model unit (Hirsimäki et al., 2006; Carki et al., 2000; Arısoy et al., 2009). To build sub-word or morpheme-based language models, text data are usually automatically segmented into sub-units based on morphological analysis and/or statistical segmentation methods. Traditionally, the segmentation methods rely on statistical models, which need reasonable amounts of annotated text data to be reliably trained. While sub-unit based language model approaches may ease the data sparsity problem compared to word-based language models, the lack of data for low-resource languages jeopardizes the training of reliable segmentation models.

In this work, we aim to improve the performance of language models for morphological rich and low-resource language with the example of Uyghur by leveraging data from a resource-rich donor language. As donor language we selected Turkish since it also uses an agglutinative morphology and shares many linguistic features with Uyghur.

To explore the impact of data from the donor language, we compared language models trained on Uyghur data only with language models trained on data from both languages, Uyghur and Turkish. Furthermore, we investigated word-based and morpheme-based language models to address the low-resource and agglutinative features of Uyghur. To study various low-resource conditions, we created different subsets of Uyghur training text data. The resulting language models are evaluated in terms of Perplexity (PPL), n-gram coverage and OOV rates.

This paper is organized as follows: in section 2 we describe the text corpora of Uyghur and Turkish. In Section 3, we introduce some common linguistic properties of Uyghur and Turkish. In Section 4, we describe the experimental set up. In Section 5, we discuss the results of our experiments.

2. Data

Uyghur is an under-resourced language with about 11 million speakers, who are mainly located in western China and Central Asia. Uyghur belongs to the Turkic language family and is closely related to Turkish. Both languages use agglutinative morphology, share features like the order of object-verb constituents, and are in parts mutually intelligible, in particular on the subject of numbers and pronouns.

The Uyghur and Turkish text data used in this study were collected by applying the GlobalPhone corpus collection procedures as described in (Schultz, 2002). As of today, the Globalphone corpus comprises of more than 450 hours of high-quality clean speech recorded from more than 2000 native speakers reading newspaper articles (Schultz et al., 2013).

2.1. Uyghur and Turkish Text Data

The Uyghur data collection, partially funded by NSF (award 1519164), comprises of news articles read by 46 speakers, as described in Abulimiti and Schultz (2020). While Uyghur is written in three different writing systems (Arabic, Roman, and Cyrillic alphabet), our corpus consistently uses Roman script.

In this work, we used the transcripts of the Uyghur ASR training data as source for language model training and the ASR development set for evaluating the language models. Table 1 summarizes the statistics of the used Uyghur text data.

	Uyghur		Turkish
	Training	Development	Training
Speakers	37	4	79
Utterances	3380	400	5482
Word tokens	60084	7902	87733

Table 1: Uyghur and Turkish text data

Turkish is used as the donor language and we use the GlobalPhone resources of the Turkish ASR training data to train the morpheme-based segmentation models and language models. The statistics of the used Turkish training text data are given in Table 1.

Since this study is meant to establish a proof-of-concept for bilingual language modeling, we focused on small amounts of Turkish data first. In future steps we plan to use larger available text corpora of Turkish, which have been collected for example within the GlobalPhone project (Carki et al., 2000).

3. Similarity of Uyghur and Turkish

3.1. Morphological Productivity

Uyghur and Turkish are both agglutinative languages, i.e. words consist of morpheme sequences (including stems and affixes) to determine their meaning, but morphemes are not altered in the process of concatenation. Typically, new words in Uyghur and Turkish are generated by adding suffixes to the end of the word. Examples of the morphological productivity are given for Uyghur and Turkish in table 2.

Uyghur words and meaning	
mektep	school
mektep-ler	schools
mektep-ler-i	of schools of third person
mektep-ler-i-de	at schools of third person
Turkish words and meaning	
iş	work
iş-çi	worker
iş-çi-ler	workers
iş-çi-ler-in	of workers

Table 2: Examples of Morphological Productivity

Uyghur and Turkish not only share a similar morphological productivity, but also have a large number of suffixes in common. We thus hope that these similarities may help to improve a morpheme-based Uyghur language model when adding morphologically segmented Turkish text data.

3.2. Mutual Intelligibility

In statistical count-based n-gram language models, every surface form of a word is modeled separately (Goodman, 2001; Tsvetkov et al., 2016). One way to improve the language model of a low-resourced language may be to make use of overlapping words from a closely-related language (Fügen et al., 2003). However, the amount of overlapping words between languages is usually not very large, even when they are closely-related. One reason is that the spelling of words may follow different writing conventions. Uyghur and Turkish share many overlapping words, e.g. "*merhaba (hello), güzel (beautiful), ölum (death), kitap (book)*", with same meanings and written form. Such overlapping words might be useful when building Uyghur language models with the help of Turkish text data. Overlapping words in Uyghur and Turkish commonly appear mostly in daily communications. In our corpus of speech read from news articles, the rate of overlapping words is thus limited. We observed 9.32% of Uyghur words in the development which appear in the Turkish training data. They corresponds to 90.60% OOV rate in the Uyghur development set.

Nevertheless, the mutual intelligibility of these two languages allows to achieving a fair amount of overlapping words. In addition, there are plenty of words, specially numbers and pronouns, which share the same meaning and similar pronunciations with only slightly different spelling. Table 3 shows some examples.

Uyghur	**IPA**	**Turkish**	**IPA**	**in English**
we	/vɛ/	ve	/vɛ/	and
ishchi	/iʃʧi/	işçi	/iʃˈʧi/	workers
üch	/yʧ/	üç	/ˈyʧ/	three
ikki	/ˈiʰʧː i/	iki	/iˈci/	two
qarar	/qɑrɑ̄r/	karar	/kɑˈrɑr/	decision
yapon	/japon/	japon	/japon/	japan

Table 3: Words with same meaning but different spelling

From many frequent words in both languages, we noticed joint spelling "patterns". For example, the graphemes in Turkish, *"ç,ş,ı"* correspond to graphemes in Uyghur *"ch, shi, i"*, respectively. After mapping the Turkish graphemes to the corresponding Uyghur graphemes, we gained more overlapping words. The words, such as *"iş (work), üç (three)"* in Turkish were mapped to *"ish, üch"*, respectively and have same spelling form as Uyghur words *"ish (work), üch (three)"* without changing the meaning.

In addition, we know that the numbers contribute to mutual intelligibility of the two languages. Therefore, numbers in Turkish spelling form are mapped to Uyghur spelling form. In this study, 30 mapping rules in total are used on Turkish data to convert the spelling form of words in Turkish to Uyghur. After applying the mapping rules, 17.08% of words in the Uyghur development set were covered by the words in the Turkish training data (82.91% OOV on

Uyghur development data). This corresponds to 7.76% absolute OOV-rate reduction compared to the Turkish data without any pre-processing. After mapping, the 100 most frequent overlapping words along with their frequencies in Uyghur and Turkish data were selected for an exemplary presentation in Figure 1.

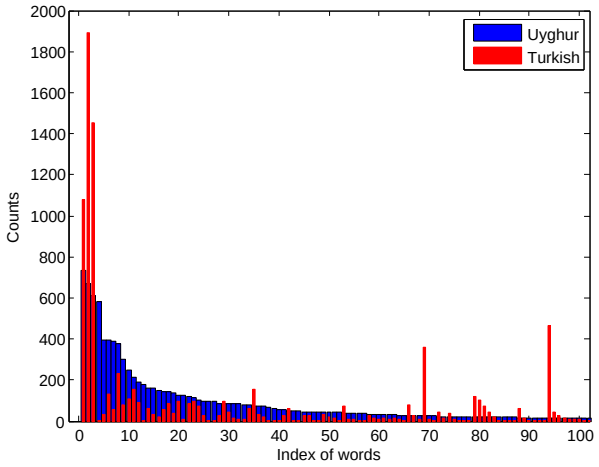

Figure 1: Counts of the 100 most frequent Uyghur words in the training data and overlap with Turkish data

4. Experiments

The main reason for using data from a donor language in a low-resource language modeling is to gain more overlapping words and get more context coverage. For low-resource and morphological rich languages, building more reliable word-based statistical language models with data from a donor language can still be challenging, due to the data sparsity problem and insufficient context coverage. In this work, we investigate how to improve the performance of language models for Uyghur using data from Turkish. We conducted two sets of language modeling experiments, word-based and morpheme-based language modeling. To explore the impact of data from the donor language, we compared language models using Uyghur data only and bilingual data (from Uyghur and Turkish). For building morpheme-based language models, morphological segmentation is done using the open source software Morfessor (Virpioja et al., 2013). It is used for unsupervised morphological segmentation of words into morpheme-like units. For training and evaluating language models, we used the SRILM toolkit (Stolcke, 2002). For word-based and morpheme-based language models, trigram models are trained by applying modified Kneser-Ney discounting (James, 2000) without cut-offs. For all the word-based language models, the words from the full Uyghur training text data (vocabulary size: 8819) is used as vocabulary. For the morpheme-based language models, this vocabulary is segmented applying the segmentation model trained with the data in the training set and then used as vocabulary of the morpheme-based language model.

To investigate various levels of low-resource conditions, we generated subsets by randomly selecting Uyghur utterances from each speaker in Uyghur training text data with varying utterance size. We collected 6 sets of Uyghur training text data for our experiment. The size of utterances, number of words and word tokens in each set are shown in Table 4. In the bilingual data experiments, these data sets are combined with pre-processed Turkish data, as discussed in Section 3.2.

Training set	Utterances	Words	Word tokens
UY_100	100	1234	1841
UY_200	200	2105	3607
UY_1k	1000	5731	17505
UY_2k	2000	7999	35249
UY_3k	3000	8783	53410
UY_all	3380	8819	60084

Table 4: Data sets used for training language models

4.1. Language Modeling on Uyghur Data Only

With the training data from those 6 sets of Uyghur data, word-based trigram language models are trained. To train morpheme-based language models for every set of Uyghur data, a segmentation model is trained with Uyghur data. Then this is used to segment Uyghur training data, Uyghur development data and the language model vocabulary. Finally, a trigram language models is built based on the segmented Uyghur training data.

4.2. Language Modeling on Bilingual Data

Since we use the same Turkish data for each set of Uyghur training data, a word-based trigram language model is trained using Turkish data and Uyghur vocabulary as mentioned above. Afterwards, for each set of Uyghur training data, one word-based trigram language model is built. For each set of Uyghur data, the best interpolation weight of Uyghur language model and Turkish language model is calculated on a held-out set. This weight is then used to interpolate the Uyghur and Turkish language models.

Morpheme-based language models for each set of Uyghur data are built with the following steps. Firstly, the Uyghur training data is merged with Turkish data and the merged data is used for training the segmentation model. After training the segmentation model, Uyghur training data, Uyghur vocabulary, Uyghur development data and Turkish data are segmented with this segmentation model. Then, morpheme-based language models are trained with segmented Uyghur data and segmented Turkish data using the segmented Uyghur vocabulary. Similar to the word-based language models, these language specific morpheme-based languages are interpolated with the best interpolation weight.

5. Evaluation

The trained word-based and morpheme-based language models for each set of Uyghur training data are evaluated on the Uyghur development set. As described in Section 4, the Uyghur development set is segmented into morpheme-like

units like Uyghur training data. Evaluation is conducted on the segmented Uyghur development data.

5.1. Word-Based Language Modeling

In Figure 2, the trigram perplexity results of word-based language models trained with Uyghur data only are compared with the interpolated language models trained with bilingual data. As can be seen, the interpolated language models outperform the Uyghur-only language models for all training set conditions. The relative improvements in terms of perplexity range from 98% to 70%. The smaller the Uyghur data in training, the higher is the relative improvement. Considering the findings in Figure 1, we assume that the amount of overlapping words in Turkish contribute to the perplexity improvements. Particularly in the small sets of Uyghur data, the overlapping word tokens of the Uyghur training data get significantly more counts, which might explain why relative reductions are higher on smaller amounts of Uyghur data.

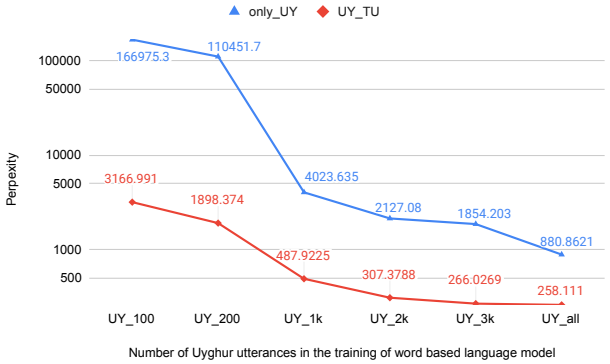

Figure 2: Perplexities of word-based language models trained with Uyghur-only and with bilingual data

Figure 3 presents the bigram coverage of language models on the Uyghur development set. With bilingual data, slightly higher bigram coverages (from 15.71% to 2.39%) are achieved relative to the language models trained with only Uyghur data. For smaller amount of Uyghur training data, the bigram coverage differences are more prominent. For the trigram coverage, we found no big difference between language models trained with Uyghur-only versus bilingual data. However, we achieved about 2% relative OOV-rate reduction on the Uyghur development set with bilingual data compared to Uyghur data only.

5.2. Morpheme-Based Language Modeling

Figure 4) shows the results on the comparisons of interpolated morpheme-based language models in terms of perplexities. As can be observed, the language model trained with bilingual data (red line) outperforms the corresponding language model trained with Uyghur data only (blue line) on all Uyghur data sets. The relative improvements in terms of perplexity range from 40.91% to 1.77%. Furthermore, for the small Uyghur training sets with 100 (UY_100) and 200 (UY_200) utterances only, the relative gains by using bilingual data are larger than for the other sets.

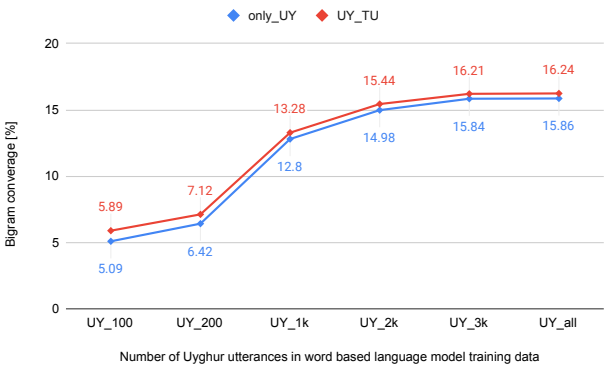

Figure 3: Bigram coverage of word-based language models on Uyghur development set

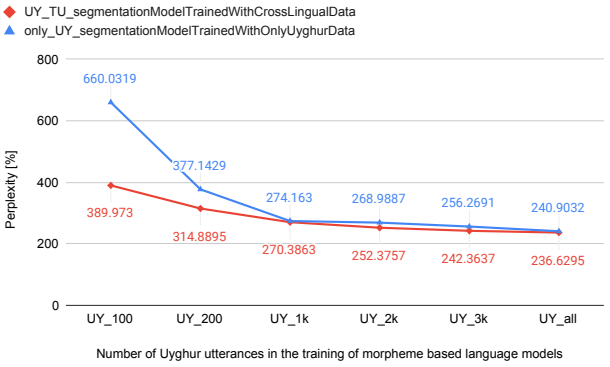

Figure 4: Perplexity of morpheme-based language models trained with only Uyghur data and bilingual data

As shown in Figure 5, the language models with bilingual data have a higher bigram/trigram coverage than the corresponding language models trained with Uyghur data only. Similar to the case of word-based language models, for the smaller data sets language models trained with bilingual data show higher relative improvement in terms of bigram/trigram coverage compared to language models trained with only Uyghur data. For example, on the UY_100 set, the language model with bilingual data has achieved 60.99% with 46.46% relative improvement in terms of bigram and trigram coverage, respectively.

As expected, morpheme-based language models result in much lower perplexities than the corresponding word-based language models. With morpheme-based data, the OOV-rate is significantly reduced compared to the word-based ones. However, regardless of the segmentation level (word- or morpheme-based), the bilingual language models outperform the Uyghur-only language models in all our experiments. Also, we observe that relative improvements are higher in terms of perplexity, n-gram coverage and OOV-rate for smaller Uyghur data sets.

Our experiments were based on the hypothesis that no morpheme-like segmentation model or morphological analysis is available for the low-resourced language. There-

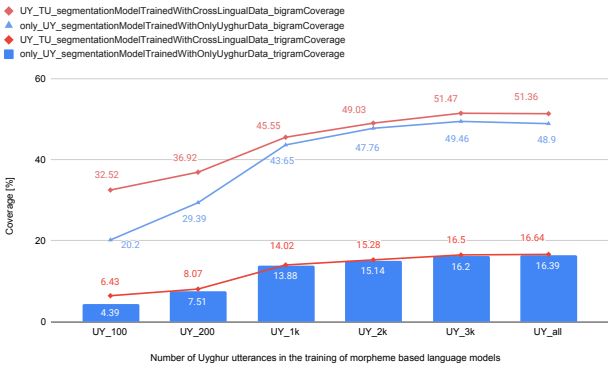

Figure 5: Bigram and trigram coverage of morpheme-based language models on the Uyghur development set

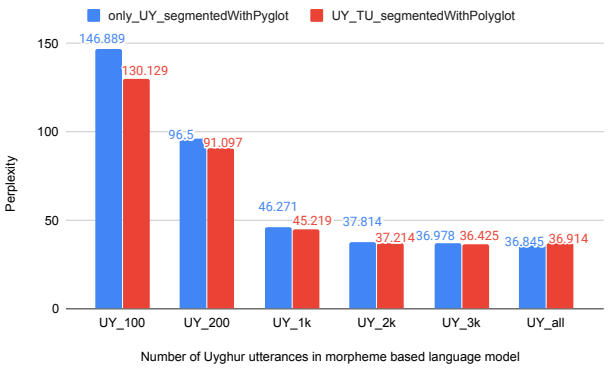

Figure 6: Morpheme-based language models using segmentation model from Polyglot

fore, for each data set, unsupervised statistical segmentation models were trained with the data from each training set. As the training data is limited, the quality of the segmentation model may be sub-optimal. Fortunately, there are software tools like Polyglot (Al-Rfou et al., 2013a; Al-Rfou et al., 2013b), which provide Morfessor models for 135 languages, including Uyghur. To explore the impact of existing and more "reliable" segmentation models, we conducted the morpheme-based language modeling experiments using the segmentation model from Polyglot.

The experiments are conducted in the same fashion as described above. The only difference is that the segmentation model from Polyglot is employed and corresponding data was segmented with that model.

Figure 6 compares the evaluation results of morpheme-based language models using Uyghur data only with bilingual data, which are segmented with Polyglot. Similar to the results from our previous experiments, the relative improvement is higher on small Uyghur data sets when bilingual data are used. On UY_100 and UY_200 set, 11.40% and 5.59% relative improvements in terms of perplexity are achieved by interpolated language models using bilingual data. However, on the sets with larger amount of Uyghur data, i.e., UY_2k and UY_3k, there is only a minor improvement with bilingual data. By the experiments using all Uyghur data (UY_all), the language model using Uyghur data only even has slightly lower perplexity. From these results, we conclude that if there is a reasonable segmentation model, language modeling with bilingual data is more suitable when only very limited data are available in the target language, for instance, under 1000 utterances. In addition, it is noticeable that the perplexity of the language models are much lower (by a factor of ca. 5) than the language models in our previous experiments.

Regarding the bigram and trigram coverage, interpolated language models achieve higher coverage than the language models trained with Uyghur data only (See Figure 7). On smaller set of Uyghur data, the improvement over using Uyghur data only is more significant. On UY_100 set, interpolated language models have relatively higher bigram (32.05%) and trigram (19.18%) coverage. Compared to the results in Figure 5, the corresponding language models have higher bigram coverage (by a factor of 1.7) and trigram coverage (by a factor of 3) in each data set, than the language models, which are trained with morpheme units segmented with the self-trained segmentation model.

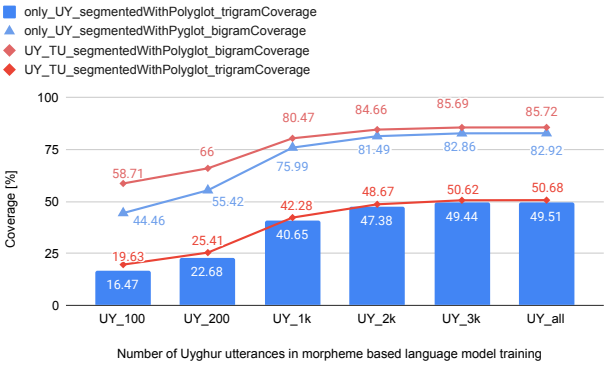

Figure 7: Bigram and trigram coverage of language models using segmentation model from Polyglot

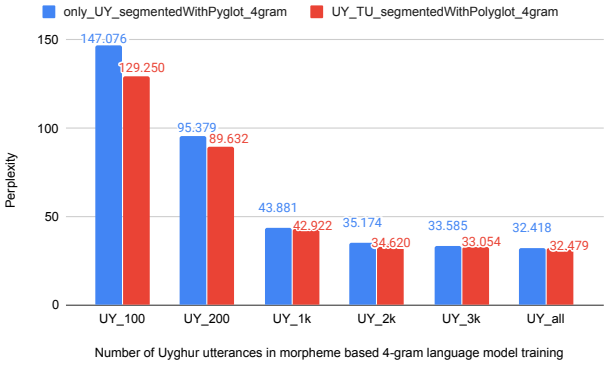

Figure 8: Perplexity of 4-gram language models using segmentation model from Polyglot

In morpheme-based language models, a word may be segmented into several morphemes, e.g. 3 morphemes. In this case, the context of the morpheme-based trigram model

may be within a word. Regarding this, we conducted the experiments using Polyglot with same fashion described above but with higher order morpheme-based language models, i.e., 4-grams. The perplexity of the morpheme-based 4-gram language models trained only with Uyghur data and with bilingual data is shown in Figure 8. Similar to the results from our previous experiments, the language models trained with bilingual data showed better performance over the language models trained only with Uyghur data. In each set of the experiments, 4-gram morpheme-based language models showed better performance in terms of perplexity over the corresponding trigram morpheme-based language models.

6. Conclusion

In this paper, we investigated word-based and morpheme-based language models for the low-resource and agglutinative language Uyghur using data from the donor language Turkish. To increase the amount of overlapping words, mapping rules are applied on the Turkish data. With this pre-processing, Turkish data achieves 7.76% of OOV-rate reduction on the Uyghur development set. Subsets of Uyghur data are generated to simulate different levels of low-resource conditions. The results indicate for both word-based and morpheme-based language models that the interpolated language model trained with bilingual data outperform Uyghur-only models in terms of perplexity, n-gram coverage and OOV-rate. Moreover, the smaller the available Uyghur data, the higher relative improvement can be achieved. Furthermore, it can be concluded that a more reliable segmentation model like Polyglot, contributes to a better morpheme-based language model regardless whether it is trained with Uyghur data only or bilingual data.

7. Bibliographical References

Abulimiti, A. and Schultz, T. (2020). Automatic Speech Recognition for Uyghur through Multilingual Acoustic Modelling. In *LREC2020*.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013a). Polyglot: Distributed Word Representations for Multilingual NLP. *arXiv preprint arXiv:1307.1662*.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013b). Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

Arısoy, E., Pellegrini, T., Sraçlar, M., and Lamel, L. (2009). Enhanced Morfessor Algorithm with Phonetic Features: Application to Turkish. In *Proceedings of SPECOM*.

Carki, K., Geutner, P., and Schultz, T. (2000). Turkish LVCSR: Towards better Speech Recognition for Agglutinative Languages. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1563–1566. IEEE.

Fügen, C., Stüker, S., Soltau, H., Metze, F., and Schultz, T. (2003). Efficient handling of multilingual language models. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 441–446. IEEE.

Goodman, J. (2001). A bit of progress in language modeling. *arXiv preprint cs/0108005*.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., and Pylkkönen, J. (2006). Unlimited Vocabulary Speech Recognition with morph Language Models applied to Finnish. *Computer Speech & Language*, 20(4):515–541.

James, F. (2000). Modified kneser-ney smoothing of n-gram models.

Schultz, T., Vu, N. T., and Schlippe, T. (2013). Globalphone: A multilingual text & speech database in 20 languages. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8126–8130. IEEE.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh International Conference on Spoken Language Processing*.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Tsvetkov, Y., Sitaram, S., Faruqui, M., Lample, G., Littell, P., Mortensen, D., Black, A. W., Levin, L., and Dyer, C. (2016). Polyglot neural language models: A case study in cross-lingual phonetic representation learning. *arXiv preprint arXiv:1605.03832*.

Virpioja, S., Smit, P., Grönroos, S.-A., Kurimo, M., et al. (2013). Morfessor 2.0: Python implementation and extensions for morfessor baseline.