Creating a Translation Matrix of the Bible's Names Across 591 Languages

Winston Wu[1]*, Nidhi Vyas[2]*, David Yarowsky[1]
[1]Johns Hopkins University, [2]Carnegie Mellon University
{wswu,yarowsky}@jhu.edu, nkvyas@cs.cmu.edu

Abstract

For many of the world's languages, the Bible is the only significant bilingual, or even monolingual, text, making it a unique training resource for tasks such as translation, named entity analysis, and transliteration. Given the Bible's small size, however, the output of standard word alignment tools can be extremely noisy, making downstream tasks difficult. In this work, we develop and release a novel resource of 1129 aligned Bible person and place names across 591 languages, which was constructed and improved using several approaches including weighted edit distance, machine-translation-based transliteration models, and affixal induction and transformation models. Our models outperform a widely used word aligner on 97% of test words, showing the particular efficacy of our approach on the impactful task of broadly multilingual named-entity alignment and translation across a remarkably large number of world languages. We further illustrate the utility of our translation matrix for the multilingual learning of name-related affixes and their semantics as well as transliteration of named entities.

Keywords: Bible, alignment, named entities, translation, transliteration

1. Introduction

In a statistical machine translation (SMT) pipeline, word alignment is important for extracting phrase translations. However, for low-resource languages with very little data, these alignments may be extremely noisy or may not exist at all. Thus, improving the quality of word alignments leads to more accurate phrase pairs, which in turn improves the quality of an SMT system (Och and Ney, 2003; Fraser and Marcu, 2006). For many low-resource languages, the Bible is the only text available, making it a valuable resource to train machine translation (MT) systems. This paper focuses on the translation and transliteration of named entities from the Bible, which are a rich resource for studying lexical borrowing (Tsvetkov and Dyer, 2016), since they are usually borrowed between languages rather than translated[1](Whitney, 1881; Moravcsik, 1978; Myers-Scotton, 2002).

Like cognates, names are often phonetically or orthographically similar across languages, which make them suited for training transliteration systems, in contrast to words which may just be translations of each other. In addition, especially for low-resource languages, names can be an invaluable source of information for learning morphemes and their semantics. However, finding their optimal translation is a challenging task, due to various reasons, including low occurrence counts and because certain names have high translation entropy or are translated into their localized proper names. In this work, we present: (1) our creation of a novel resource of 1129 English Bible named entities aligned across 591 languages; (2) novel methods to produce and clean the resource; (3) challenges and findings in this process; and (4) potential use cases of our resource. Our Bible names translation matrix is available at `github.com/wswu/trabina`. We believe this resource will be of great linguistic importance in studying low resource languages and will be applicable in several areas such as transliteration and morphological analysis of named entities.

2. Related Work

Due to low frequency words, word alignments can suffer from misalignments, which in turn can be detrimental to downstream tasks like MT. Previously, researchers have worked on improving word alignment to improve MT using a variety of approaches: combining hypotheses generated from bridge languages (Kumar et al., 2007), using semantic relationships (Songyot and Chiang, 2014), prior distributions (Mermer and Saraclar, 2011; Vaswani et al., 2012), discriminative alignment models (Moore, 2005; Taskar et al., 2005; Riesa and Marcu, 2010) and part-of-speech (POS) tagging (Lee et al., 2006; Sanchis and Sánchez, 2008). Our work uses the assumption that names are often orthographically or phonetically similar across languages. We use MT as an intermediate step to generate hypotheses for citation form alignments. This is akin to training phrase-based MT systems for transliteration (Song et al., 2009; Jia et al., 2009; Dasigi and Diab, 2011). (Dasigi and Diab, 2011) used a character-based Moses system and post-edited the output using linguistic rules, which is similar to our approach of using MT and applying transformation rules. Our approach is unique in that we use cross-language joint models of variant/latent forms to expand and refine the candidate space of citation forms.

3. Translation Matrix

The primary contribution of this paper is a translation matrix of 1129 English names aligned and translated into 591 languages. We produced around 14,000 alignments, providing better coverage than Wikipedia for several names.[2] On average, a name contains realizations in 52% of all languages.[3] The names in this matrix were extracted from the parallel Bibles corpus (Mayer and Cysouw, 2014), in which verses

*Denotes equal contribution.

[1] The opposite case is also very interesting, e.g. in many languages, *Caesar* is translated as *emperor* rather than transliterated.

[2] For example, in Wikipedia, *Egypt* is translated (including Romanization) into 55 languages, whereas our resource contains translations of Egypt in the majority of 591 languages.

[3] Names that occur frequently, like *Jesus*, are covered in almost all 591 languages, while uncommon names or name variants, like *Antiochia* (variant of *Antioch*), appear in fewer languages.

English	Yipma	Hanga	Koongo	Hanunoo	cnw*	Maltese	Latvian	gug*
jesus	jizaazai	yeesu	yesu	hisus	jesuh	ġesú	jēzus	jesús
christ	kɨraazɨ	kristu	klisto	kiristu	khrih	kristu	kristus	cristo
israel	yɨzɨrelɨ	juusi	isaeli	israil	israel	iżrael	izraēli	israel
david	devitɨ	dawuda	davidi	dabid	david	david	dāvida	david
paul	polɨ	pool	pawulu	pablu	paul	pawlu	pāvils	pablo
peter	pitai	piita	petelo	pidru	piter	pietru	pēteri	pedro
egypt	yɨzɨpɨ	yijipi	ngipiti	ihiptu	izip	eġittu	ēģipti	egípto
jerusalem	jeruzaalemɨ	jiruusilim	yelusalemi	hirusalim	jerusalem	ġerusalemm	jeruzalemē	jerusalén

Table 1: Example translation matrix of Bible named entities (*cnw = Chin Ngwan, *gug = Paraguayan Guaraní)

are aligned across all languages. Each cell in the translation matrix contains the best guess of the citation form of the English name in the target language. This form is the consensus of four different methods, which are described in the following sections. An excerpt of the name translation matrix is shown in Table 1.

4. Improving Named Entity Alignment

The source data from Mayer and Cysouw (2014) contains 24 English editions of the Bible. For 591 target language bibles, we word aligned each verse with each English Bible verse using the Berkeley Aligner (Liang et al., 2006) and performed POS tagging to extracted proper nouns from these alignments. A total of 1129 English named entities were extracted. For each English name, we considered multiple citation hypotheses in all target languages from the following four approaches.

4.1. Most frequent alignment

For every target language, the baseline hypothesis for a name's translation is the most frequent alignment obtained from the aligner. These initial hypotheses contained several alignment problems which we broadly classify into three categories: (1) incorrect alignment, (2) missing alignment, and (3) non-base form alignment.[4] Examples of these alignment errors are presented in Table 2. We improve on these initial alignments by tackling each of these problems in turn.

4.2. Distance-based approach

Incorrect alignments (Issue 1) are obviously problematic. Using the assumption that names are borrowed very frequently across languages and undergo minimal orthographic change, we employed an edit distance based approach to produce a hypothesis. If we let A be the top 5 most frequent alignments within a language, and T be the top 5 most frequent alignments combined across all languages,[5] the distance-based approach selects the name $h_{dist} \in A$ with the smallest Levenshtein distance (Levenshtein, 1966) to any word in T. This approach resolved many incorrect alignments.

For example, in the Manikion bible, the English name *Boas* was aligned 21 times to four words: Obed (9), Boas (7), eici (4), and Nahson (1). Clearly, the original alignment (Obed) is not the best translation. Taking the top 5 most common alignments for *Boas* across all languages (Booz, Boas, Boaz, —[6], and Boasi), we calculated the edit distance between each pair of words. The Manikion name with the minimum distance to any of the language-independent names (Boas) is a better translation than the original baseline alignment (Obed).

	Booz	Boas	Boaz	–	Boasi
Obed	4	4	4	4	5
Boas	2	**0**	1	4	1
eici	4	4	4	4	4
Nahson	5	5	6	5	5

4.3. Producing hypotheses with MT

Missing alignments (Issue 2) are common for low-frequency words, due to a weak alignment signal, or in certain languages, because these words may not exist at all in the Bible[7]. Such issues cannot be overcome using the distance-based approach, which selects a hypothesis from existing alignments. Thus, we employ character-based machine translation to suggest possible foreign names to which an English name is aligned.

Using the translation matrix with the distance-based hypotheses as bitext, we generate multiple hypotheses for plausible translations/transliterations by training ten MT systems for a single target language. The source languages were: four pivot languages (English, French, Spanish, Italian) and the six nearest languages in an ordering based on the language tree in Ethnologue. The pivot languages were chosen for their near-complete coverage over the 1129 names, and we utilized the six nearest languages due to the potential for names to be orthographically similar in related languages.

For each language pair, treating foreign-English name pairs as bitext, we split the data in half and trained two systems A and B, such that system A decodes the data that system B was trained on, and vice versa; this was done to ensure that the test set was never seen by the system performing decoding. We used a standard Moses (Koehn et al., 2007) setup with 5-gram KenLM (Heafield, 2011) language model and MERT (Och, 2003) for tuning. Each system generated a unique 200-best list of hypotheses.

[4] Throughout this paper, *italicized* names in tables refer to incorrect alignments/hypotheses.

[5] These five names represent a language-independent consensus; the intuition is that regardless of the language, the best translation of a name should be similar if not equal to one of these top five. We selected a threshold of five names because we found that this covered the major realization variants of a name.

[6] Denotes missing alignment, which in this case is frequent enough to make it into the top 5.

[7] For example, a language may not have a translation of the Old Testament, so names appearing only in the Old Testament will have missing alignments.

(a) Issue 1: Incorrect alignments

English	Foreign	Lang
Boaz	*Obedarinchichitam*	cbu
Boaz	*Obed*	mnx
Obed	*Jeseyrinchichitam*	cbu
David	*Luwiy*	agu
Eliezer	*Yesua*	mnx
Julia	*Pirorogasomɨ*	aak

(b) Issue 2: Missing alignments

English	For.	Lang
Mnason	—	cbu
Phoenix	—	cbu
Dionysius	—	mnx
Illyricum	—	mnx
Ephphatha	—	aak
Colossae	—	aak

(c) Issue 3: Non-lemma alignments

English	Foreign	Lang
David	*Dapiyarin*	cbu
Eliezer	*Eriesaorɨnɨ*	aak
Eliezer	*Elíyaserarinchichitam*	cbu
Aram	*Ramaho*	mcq

Table 2: MLE alignment problems: cbu = Candoshi, mnx = Manikion, aak = Muak Sa-aak, mcq = Ese, agu = Awakateko

4.3.1. Filtering and scoring hypotheses

The hypotheses generated by the MT approach were not necessarily valid names. To rectify this, we combined the n-best lists of all ten systems, filtered out hypotheses that did not occur in the target language's bible, and rescored the remaining hypotheses with a weighted combination of four features:

Fraction of observed to total count: The observed verse count of a hypothesis is a measure of how often it is aligned to the corresponding English name. It was calculated as follows: for each Bible verse where the English name occurred, we looked into the corresponding foreign Bible verse and within a window of ±3 verses[8], and incremented a count if the hypothesis was present. The total count is the number of times the hypothesis appeared in the entire foreign Bible.

Model score: This feature incorporates the MT decoder's score, which we observed to generally fall within the range -5 to 5. We normalized it to bring it into the range [0, 1].

Closeness to expected English count: This feature gave weight to the actual coverage of a hypothesis within a given Bible. For a good hypothesis, we expect its observed verse count c_v to be close to the corresponding expected count in English. Since there are 24 English bibles, the expected English count c_e is the total English count ÷ 24. We defined closeness as $\frac{-(c_v - c_e)^2}{c_e^2}$ if $c_v < c_e$ and c_e/c_v otherwise. The rationale for using a piece-wise function is that if the observed verse count is less than the expected count, then it might be due to inflected forms being aligned to the word. However, if the observed verse count is higher than the expected count, then it is likely not the correct alignment.

Matches the gold: 1 or 0 if the hypothesis matches the gold name h_{dist}.

We manually tuned the weights to $[0.4, 0.4, 0.05, 0.05]$ for the above features, respectively. Higher scores indicate better hypotheses. Unlike the baseline aligner and distance-based approaches, the MT approach can generate multiple hypotheses for a single English name and can produce plausible translations for missing alignments (see Table 3).

English	Aligner	MT Hypotheses	Lang
Ephphatha	—	Epata	aak
Colossae	—	Korosi	aak
Colossae	—	Kolose	mnx
Dionysius	—	Dionisius	mnx
Illyricum	—	Ilirikum	mnx
Mahalaleel	—	Mahalelel	mnx
Dalmatia	—	*Tármatiyap*	cbu
Phoenix	—	Finíase	cbu
Sergius	—	Sírjiyu	cbu
Mnason	—	*Nasónap*; Nasón	cbu

Table 3: Hypotheses generated by MT approach, filling in missing alignments. aak = Muak Sa-aak, cbu = Candoshi, mnx = Manikion

Pairs in C-Set	LCS	T-Rule
Yose, Yose	Yose	∅ ↔ ∅
Yose, Yose'nin	Yose	∅ ↔ 'nin
Yose, Yusuf	Y-s-	o-e ↔ u-uf
Yose'nin, Yusuf	Y-s-	o-e'nin ↔ u-uf
Beytanya, Beytanya	Beytanya	∅ ↔ ∅
Beytanya, Beytanya'ya	Beytanya	∅ ↔ 'ya
Beytanya, Beytanya'dan	Beytanya	∅ ↔ 'dan
Beytanya'ya, Beytanya'dan	Beytanya	'ya ↔ 'dan

Affix	Freq
∅	7
'dan	2
'ya	2
-u-uf	2
'nin	1
o-e	1
o-e'nin	1

Table 4: Transformation rules extracted for Turkish hypotheses of the English names *Joses* and *Bethany*, and their affix frequency.

4.4. Transformation rules

Non-base form alignments (Issue 3) are common and occur when an English name, which does not mark case, is aligned to an inflected foreign name. String transformation rules were employed to recover the best candidate h_{tr} for the citation form of the English name in the target language from a set of candidate hypotheses generated by the previous approaches (Sections 4.1. to 4.3.).

Candidate hypotheses set (C-Set) This set consists of a total of 6 hypotheses (or less if any hypothesis from the previous iterations was missing):

1. Baseline alignment
2. Hypothesis from distance-based approach
3. Most frequent 1-best hypothesis from each MT system
4. Most frequent hypothesis from combined hypotheses
5. Best scoring 1-best hypothesis from each system

[8] This compensated for possible incorrect verse alignments in the data that we started with.

English	Aligner	T-Rule	Lang
Boaz	*Bowasomɨ*	Bowaso	aak
Eliezer	*Eriesaorɨnɨ*	Ereasao	aak
Esau	*Isomɨ*	Iso	aak
Israel	*Isɨrerɨyí*	*Isɨrene*	aak
David	*Dapiyarini*	Tapít	cbu
Eliezer	*Elíyaserarinchichitam*	Elíyas	cbu
Rama	*Aramho*	Ramo	mcq

Table 5: String transformation rules recover the lemma form of inflected alignments. mcq = Ese

6. 2nd best scoring 1-best hypothesis from each system

Transformation Rules (T-Rule) We define T-Rules as language-specific string transformation rules that change one word form to another. For example, the T-Rule $\{\emptyset \leftrightarrow s\}$ can change a singular to a plural word in English (e.g. *Cat* ↔ *Cats*). For an English name's C-Set, we construct a T-Rule for every pair of words in the C-Set by removing the longest common subsequence between them (see Table 5). After obtaining rules for all English names in a given target language, we combined and sorted them by their frequency (Table 4). The hypothesis in the C-Set with the maximum underlying transformation frequency in the combined list of affixes was selected as the best translation hypothesis h_{tr}. For example (Section 4.4. and table 4), the empty affix is the most frequent affix, so this approach takes this to be the affix for the citation form of a name and thus selects *Yose* and *Beytanya* as the new base translations for Joses and Bethany, respectively. This method leverages the globally distributed information of word transformations in the target language to select the citation form. Table 5 shows examples of such cases.

5. Evaluation

To evaluate each alignment-improving approach, we first acquired a manually annotated test set of 30 randomly-chosen names across 591 languages, for a total of 17,730 examples. Annotators were asked to verify that the foreign best was indeed the best translation of that name. If not, they were to replace it with either a name from the alternatives (hypotheses from the various methods described above) or a blank if they deemed that none of the choices were a good translation. We used three annotators, and each test example was examined by two annotators. The data was shown in a tabular format, which facilitated inspection as well as provided stimuli from nearby languages. If there was a disagreement, we randomly chose one of the words to be the gold. Although the annotators do not know all 591 languages, the average inter-annotator agreement over the entire test set was 0.92, indicating that they have a good intuition of what names should look like even in languages that they are not familiar with.[9]

	Align	Dist	MT	TR	Maj.	WC
Average	.682	.762	.805	.499	.778	.779

Table 6: Average 1-best accuracy on test words. TR = transformation rules, Maj. = majority vote, WC = weighted combination

English	Lng	BA	Distance	MT	TR	Consensus
Pharaoh	ctu	*egipto*	faraón	faraón	faraón	faraón
Tobias	por	—	tobias	tobias	tobias	tobias
Caesar	bzj	seeza	*roam*	seeza	*koam*	seeza
Phoenix	bcw	fenikəsə	fenikəsə	fenikəsə	feniki	fenikəsə
Pyrrhus	gbi	*sopater*	*sopater*	*sopater*	pirus	*sopater*
Zion	msm	sion	sion	sion	*siam*	sion
Claudia	agg	krodia	krodia	krodia	*kardia*	krodia
Claudia	cbu	*linu*	*linu*	*linu*	*linu*	*linu*
Pyrrhus	lac	*sópater*	*sópater*	*sópater*	*berea*	*sópater*

Table 7: Examples of best hypotheses from each system

6. Results

We evaluated the accuracy of the four methods on predicting the correct test word, in addition to system combination via majority voting and weighted combination (Table 6). Note that these methods are not independent of each other, since the MT approach builds on the results of the distance-based approach, and the transformation rules approach build upon the previous two. Results are across all 591 languages, and the size column indicates the number of non-empty gold words evaluated against. The weighted combination was done by taking a weighted consensus of the top hypothesis by each of the four approaches (Aligner, Distance, MT, and T-Rule), weighted by the average performance of each approach (.683, .763, .805, and .449, respectively).

On average, we find that the distance based approach and the MT approach perform comparably, showing large improvements over the baseline. We see that using a simple majority consensus to combine the outputs of the four methods, while not as good as MT alone on average, obtains the best performance on 12 of the test words, in contrast to 10 for MT. Our methods effectively generated citation forms for words that were not seen before by the aligner, which suggests that these methods can be effective in generating new vocabulary for low resource languages.

Some examples from each approach are presented in Table 7. Pharaoh and Tobias represent typical cases where our approaches improve upon the aligner baseline. The next few names show examples of disagreement between the different approaches. Claudia and Pyrrhus exemplify cases where none of the approaches chose the correct translation, largely due to high co-occurrence count of these incorrect names and their English counterparts, which caused misalignments.[10] On test words with low frequency across languages (Havilah, Jarmuth, Shishak, Mordecai and To-

[9] Obviously annotators do not know all 591 languages. However, the large majority of names are related across languages (e.g. Mesir/Masr/... or Eiypta/Ehipto/...) and typically undergo systematic sound/orthography shifts between languages. Because of this, annotators can readily pick out the correct name with high interannotator agreement, especially after having seen the name's realizations in related languages.

[10] In the Bible, Claudia and Linus occur frequently together, as do Pyrrhus and Sopater (Pyrrhus' son) and Berea (the city they were from).

bias), our methods show 13–72% improvements over the baseline, suggesting that our approaches can overcome the problem of limited size of the data set.

In total, we found that the distance-based approach changed 1935 alignments (∼98/name), the MT approach changed 5516 existing alignments (∼280/name) and generated 3170 missing alignments (∼161/name), and the transformation rule approach changed 9053 alignments (∼460/name).

Further examination of the results reveals variations in the translation of certain names, which is linguistically interesting. Some notable examples include Caesar, which is split across languages whether to be pronounced with a hard *c* (e.g. Kaiser, Keizer, etc.), as in Classical Latin, or a softer *s* (e.g. Sisa) or *ts* (e.g. Czar/Tsar). In addition, Caesar is often translated as “Emperor” or “King” rather than transliterated. Likewise, Sheol is more often translated as the language’s word for “Hades” or “Hell” than transliterated. This phenomenon is likely attributed to cultural influences.[11]

7. Applications

We present preliminary investigations into two potential use cases of our resource.

7.1. Named Entity Morphology

In the process of aligning a morphologically rich language to English, the aligner encounters many morphological variants of the same name. For example, the English name *David* is best aligned to the lemma form *Depito* in the Ankave language. However, *David* is also aligned to words of the form *Depito* + some affixes. By examining the morphological variants of *Depito*, we gather the following transformation rules:

Depito	$\emptyset \leftrightarrow \emptyset$
Depitoyá	$\emptyset \leftrightarrow$ **-yá**
Depitomɨ	$\emptyset \leftrightarrow$ -mɨ
Depitorɨnɨ	$\emptyset \leftrightarrow$ -rɨnɨ

A rule’s right hand side can be considered an affix that denotes some aspect of morphology. Then, we can discover this affix’s semantics by modeling the context surrounding the English word to which the morphological variant was aligned. The intuition is that inflected foreign forms of proper names should co-occur often with English prepositions and thus is correlated with a semantic case; or with preceding conjunctions, which can indicate plurality. By performing this process over all names in a language, we can determine the meaning of a given affix. For example, Ankave words ending with *-yá* occur frequently with the English preposition *of*, so *-yá* is very likely a marker for the genitive case. This process is especially applicable for identify the meaning of morphological affixes in low-resource languages where a grammar of the language may not exist and would be time-consuming to create by hand.

7.2. Transliteration

The Bible name translation matrix is naturally suited for training transliteration systems. By treating the weighted consensus names in our translation matrix as bitext, we trained character-based Moses SMT systems in a 80-10-10 train-dev-test split on a random subset of 40 languages, with the target language being English. We use a standard setup with a 4-gram language model, tuning with MERT, and no distortion to prevent reordering.

We compare against a simple baseline, Unidecode[12], which provides context-free Unicode to ASCII character mappings. While this is a naive baseline, it is reasonable for many low-resource languages for which this is perhaps the only form of transliteration available. Even on the order of several hundred training examples, underscoring the low-resource nature, the MT systems trained on this data transliterated on average much better than the baseline (Figure 1).

Figure 1 compares the accuracy of the baseline versus a Moses-based transliterator. The languages are represented by their ISO 639-3 language codes. Overall, the average accuracy of the baseline was 0.17, compared to 0.23 for Moses. Note that for the two highest scoring languages, ifb (Ifugao, a Malayo-Polynesian language) and bvr (Burarra, an Australian Aboriginal language), the Unidecode baseline performed better than Moses. For these two languages, most of the source and target words were identical. In such cases, Moses would just learn character identity mappings, and we suspect that the language model was biasing the system away from the correct answer. For example, the Ifugao-English system incorrectly transliterated *Amminadab* as *Aminadab*, whereas passing the source through unchanged would have achieved a higher accuracy. More investigation is necessary to determine the role of the language model in transliteration, especially of low-resource languages.

A followup on this work is presented in Wu and Yarowsky (2018), who compare the performance of several methods of transliteration, including phrase-based and neural machine translation, on our Bible names dataset.

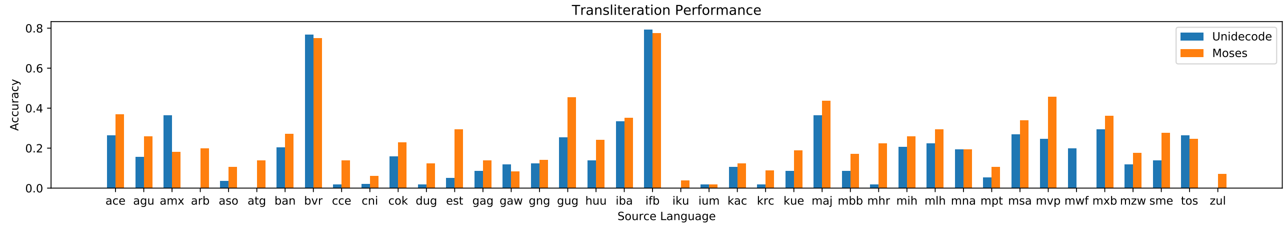

Figure 1: Comparing the performance of a baseline transliterator with a SMT-based transliterator trained on our translation matrix

[11] For example, Caesar was originally just a name but eventually became a title for the Roman emperor.

[12] `https://pypi.python.org/pypi/Unidecode`

8. Conclusion

We have developed and presented a novel and impactful dataset of 1129 person and place names in the Bible aligned over 591 languages. We have presented and empirically contrasted several techniques, including distance-based metrics, machine-translation-based transliteration models, and affixal transformation rules, for iteratively refining alignments. Our improved alignments outperform baseline alignments from a widely-used word alignment software in 97% of words in the test set. We release our Bible names translation matrix dataset, which we believe will be of value to researchers looking to build transliteration systems or other applications for low resource languages, for which the Bible may be the significant available bilingual, or even monolingual, text.

9. Acknowledgments

This work was supported in part by the DARPA LORELEI program. The findings, conclusions, and opinions found in this work are those of the authors and do not necessarily represent the views of the funding agency.

10. Bibliographical References

Dasigi, P. and Diab, M., (2011). *Proceedings of the 3rd Named Entities Workshop (NEWS 2011)*, chapter Named Entity Transliteration Generation Leveraging Statistical Machine Translation Technology, pages 106–111. Asian Federation of Natural Language Processing.

Fraser, A. and Marcu, D. (2006). Semi-supervised training for statistical word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 769–776. Association for Computational Linguistics.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Jia, Y., Zhu, D., and Yu, S., (2009). *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, chapter A Noisy Channel Model for Grapheme-based Machine Transliteration, pages 88–91. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Kumar, S., J. Och, F., and Macherey, W. (2007). Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Lee, J., Lee, D., and Lee, G. G. (2006). Improving phrase-based korean-english statistical machine translation. In *INTERSPEECH*.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710, February.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Mermer, C. and Saraclar, M. (2011). Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 182–187. Association for Computational Linguistics.

Moore, R. C. (2005). A discriminative framework for bilingual word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Moravcsik, E. (1978). Language contact. *Universals of human language*, 1:93–122.

Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press on Demand.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, Volume 29, Number 1, March 2003*.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Riesa, J. and Marcu, D. (2010). Hierarchical search for word alignment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 157–166. Association for Computational Linguistics.

Sanchis, G. and Sánchez, J. A. (2008). Vocabulary extension via POS information for SMT. *Mixing Approaches to Machine Translation*.

Song, Y., Kit, C., and Chen, X., (2009). *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, chapter Transliteration of Name Entity via Improved Statistical Translation on Character Sequences, pages 57–60. Association for Computational Linguistics.

Songyot, T. and Chiang, D. (2014). Improving word alignment using word similarity. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1840–1845. Association for Computational Linguistics.

Taskar, B., Simon, L.-J., and Dan, K. (2005). A discriminative matching approach to word alignment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Tsvetkov, Y. and Dyer, C. (2016). Cross-lingual bridges with models of lexical borrowing. *J. Artif. Intell. Res.(JAIR)*, 55:63–93.

Vaswani, A., Huang, L., and Chiang, D. (2012). Smaller alignment models for better translations: Unsupervised word alignment with the l0-norm. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–319. Association for Computational Linguistics.

Whitney, W. D. (1881). On mixture in language. *Transactions of the American Philological Association (1869-1896)*, 12:5–26.

Wu, W. and Yarowsky, D. (2018). A comparative study of extremely low-resource transliteration of the world's languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resources Association (ELRA).