Character Alignment in Morphologically Complex Translation Sets for Related Languages

Michael Gasser
Indiana University
gasser@indiana.edu

Nazareth Amlesom Kifle
Østfold University College
nazareth.a.kifle@hiof.no

Binyam Ephrem
Addis Ababa University
binephrem@gmail.com

Abstract

For languages with complex morphology, word-to-word translation is a task with various potential applications, for example, in information retrieval, language instruction, and dictionary creation, as well as in machine translation. In this paper, we confine ourselves to the subtask of character alignment for the particular case of families of related languages with very few resources for most or all members. There are many such families; we focus on the subgroup of Semitic languages spoken in Ethiopia and Eritrea. We begin with an adaptation of the familiar alignment algorithms behind statistical machine translation, modifying them as appropriate for our task. We show how character alignment can reveal morphological, phonological, and orthographic correspondences among related languages.

1 Background

Languages such as Turkish, Arabic, and Swahili are characterized by a wide variety of productive ways in which morphemes are combined to form words. In these *morphologically complex languages* (MCLs) —and there are thousands of them— many more specific word forms are possible than in languages such as English and Chinese. This very high type-to-token ratio results in a severe data sparsity problem, leading the field of machine translation (MT) to begin to take morphology seriously once such languages were considered (El-Kahlout et al., 2019). As far as we know, however, no one has investigated the translation of individual words between MCLs. Why would such a task be worth pursuing? First, in MCLs, words may correspond to entire phrases in languages such as English or Chinese. Thus when translating from, say, Arabic to Turkish, the capacity to translate individual nouns or verbs, segmented into their constituent morphemes, could play a significant role. We suspect this will be especially true for translation between closely related MCLs. In such cases, syntactic differences may be limited, with word order and long-distance dependencies largely preserved, while the within-word differences may remain challenging. Second, focusing on word translation pairs may reveal other potentially useful information, especially for related languages, such as language distance or correspondences between characters or phonemes. Finally, a system that translates complex words can support applications such as cross-linguistic information retrieval, dictionary creation, and language instruction.

In this paper, we consider the more general problem of translation among a *set* of words in related MCLs. As an illustration, see the words in column (b) of Figure1, each a translation of 'I was listening' in one of four closely related Romance languages of the Iberian Peninsula. By examining multiple related languages, we eventually hope to be able to capitalize on inter-relationships among the languages. We are specifically interested in the word translation task for languages with few computational resources.

Our task is related to other work in MT. We could, for example, approach the translation of pairs of words within a set using one of the state-of-the-art neural sequence-to-sequence methods, which have also been shown to be applicable to other morphological tasks (Cotterell et al., 2018; McCarthy et al., 2019). However, we may be just as interested in the details of what is learned — where the boundaries between

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 47–56
Barcelona, Spain (Online), December 13, 2020

morphemes are, what morphophonological rules apply — as we are in the performance itself, and it may be difficult or impossible to extract this knowledge from a neural system. For now, we will look for inspiration in earlier MT work, specifically variants of statistical MT (SMT) and leave comparison with neural approaches to future work.

A further related area of work, which actually resembles our task more closely, concerns morphological paradigms (Erdmann et al., 2020; Jin et al., 2020; Cotterell et al., 2018; McCarthy et al., 2019). A partial example of a Spanish paradigm appears in column (a) of Figure 1. Of particular interest to us is the "paradigm cell filling" problem, that of learning the inflections that comprise a paradigm given all forms for a set of lemmas (Durrett and DeNero, 2013).

If we compare word translations in a set of related languages with cells in a morphological paradigm within one language, the relationship between our task and the paradigm cell filling problem becomes clear. However, this is only the case for the translation of words with similar roots. Figure 1 illustrates this point with an example paradigm from Spanish and two example translation sets from four Romance languages.[1] When the roots are related, as in the translation set in column (b), the set resembles a paradigm; when they aren't (or aren't obviously), as in the translation set in column (c), only the affixes resemble one another. We will be focusing on the case where roots are similar; thus we will need to address the question of how to distinguish such sets from those like the one in column (c).

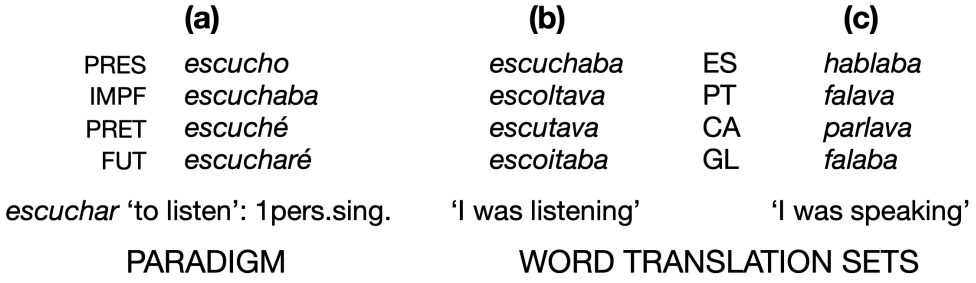


Figure 1: Paradigm (Spanish); word translation sets (Spanish, Portuguese, Catalan, Galician)

Our longer-term goal is translation: given a word in one language, generate a plausible word in one or more of the other related training languages. In this paper, however, we focus on the subtask that Durrett and Denero (2013) begin with in their paradigm cell filling research, that of aligning the characters of the words in a translation set.

The rest of the paper is organized as follows. In Section 2, after an introduction to the Ethio-Eritrean Semitic (EES) languages, we define our task in more detail, looking specifically at possible types of morphological relationships between related languages, with focus on the EES languages. Next, in Section 3, we discuss how some cross-linguistic morphological relationships can be captured in terms of character alignment and show how character alignment within word translation pairs in closely related languages can be be constrained. Next, in Section 4, we outline our approach to character alignment within word translation sets, in particular, how we adapt SMT word alignment methods to handle character-level alignment. Then, in Section 5, we describe our experiments applying the method to data from the Semitic languages. In Section 5.4, we discuss results, both in terms of alignment performance and in terms of useful information that the alignments provide. Finally, in Section 6, we conclude, looking forward to the next steps in the long-term project of the translation of morphologically complex words.

2 Semitic languages of Ethiopia and Eritrea

The languages of the world differ greatly with respect to how much morphological elaboration they make available to speakers (Stump, 2017). What is interesting for our purposes (and those of this workshop) is the apparent tendency for languages within a closely related family to *agree* in morphological complexity, though not necessarily in the morphological details.

[1] Without the benefit of context, there is a much greater potential for translation ambiguity for words than for sentences. Here and elsewhere in this paper, we assume that "translation of a word" refers to a translation in *some possible context*.

In this paper we focus on one of these families, the Ethio-Eritrean Semitic (EES) languages, consisting of the subgroup of Semitic languages spoken in Ethiopia and Eritrea (Demeke, 2001). There is no general agreement on the number of EES languages, but, judged by mutual intelligibility, there are between 10 and 15. Of these, two, Tigrinya (hereafter, Ti) and Tigre (hereafter, Te), are spoken in Eritrea, by roughly 85% of the population, and the remainder, in addition to Ti, are spoken in Ethiopia, as native languages of roughly 38% of the population. These include Amharic (hereafter Am), the official language of the Federal Government of Ethiopia. The EES languages are divided into northern and southern subgroups, with Ge'ez, Ti, and Te in the northern subgroup, Am and the remaining languages in the southern subgroup.

2.1 Morphology

The languages exhibit many features common to other Semitic languages, especially very complex verbal morphology, including subject agreement suffixes and prefixes, object agreement suffixes, and the templatic morphology that Semitic languages are famous for. Verb roots consist of sequences of consonants, organized in a number of root classes. Verb stems consist of patterns of particular vowel and consonant gemination that are overlaid on the consonantal roots. For an Am, Ti, and Te example, see Figure 3a.

2.2 Orthography

The EES languages are written using the Ge'ez writing system, an abugida system in which each character represents either a consonant-vowel sequence (the onset and nucleus of a syllable) or a single consonant (the coda of a syllable). The characters are usually presented in a table, in which each row contains the characters for a single consonant. Figure 2 shows three rows from this table, the symbols representing the consonants /*l*/, /*b*/, and /*g*/ followed by vowels or alone. In addition to their realization as a single consonant, the characters in the sixth column can also represent the consonant followed by the epenthetic vowel [*ɨ*].[2] For example, the Am word *bəlɨbb* 'by heart' is written በልብ. Of the two sixth-column characters, the first, ል, is pronounced with a following vowel, while the second, ብ, is not. This example also illustrates the only significant deviation of the orthography from the phonological form of a word: consonant gemination (length) is not indicated.

	ə	u	i	a	e	—	o
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ

Figure 2: Three rows from the table of Ge'ez characters.

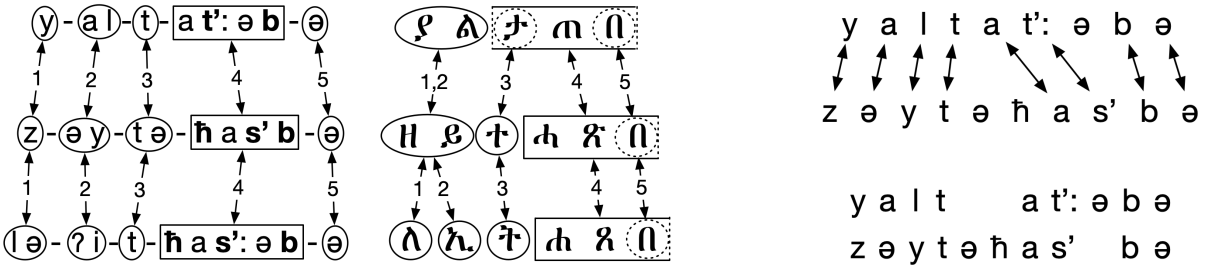
In our experiments we work with both the *orthographic* representations of words and a *phonemic* transcription in which gemination is indicated (marked with ":"). Because the orthography obscures the boundaries between consonants and vowels, it may also obscure the boundaries between morphemes. An example is shown in Figure 3a. Circles surround affixes; rectangles surround stems. In multiple cases, morpheme boundaries are between consonants and vowels, so a single Ge'ez character spans two morphemes. For example, the final character in each word, በ /*bə*/, includes the last stem consonant /*b*/ and the perfective 3sm suffix /*ə*/. We include gemination in our phonemic representation because of its role in EES verb morphology as well as its importance for speech synthesis in the languages.

2.3 Previous work

Like most African languages, EES languages suffer from a lack of resources. With the exception of Am and Ti (Abate et al., 2018), corpora in the languages are very limited, in some cases non-existent. However, we do have the advantage of bilingual root lexica (Am-Ti, Ti-Te) and dedicated morphological analyzers and generators for three of the languages, Am, Ti, and Te, within the HornMorpho project (Gasser, 2011).[3] HornMorpho consists of weighted finite-state transducers constructed by hand. Using HornMorpho and our bilingual root lexica, we can generate data for the morphological translation task.

[2] Linguists disagree on whether this vowel represents a phoneme, and its status may vary between the languages.

[3] Te has been added to HornMorpho for the purpose of this project.



(a) Morpheme alignment, phonemic and orthographic. 1–RELATIVE, 2–NEGATION, 3–PASSIVE, 4–STEM(*wash*), 5–PERFECTIVE, SBJ=3RD.MAS.SG.

(b) Monotonic character alignment (Am, Ti only), diagrammed in two ways.

Figure 3: Alignment within an Am-Ti-Te translation set. Gloss: 'which was not washed.'

3 Morphological alignment and character alignment

It is beyond the scope of this paper to discuss morphological complexity (Stump, 2017). We base our consideration of morphology in the remainder of the paper on the structure of verbs in EES languages and on general notions familiar from basic morphological research. We assume that for each word in a MCL, there is a stem, based on a root; zero or more affixes; and possibly instances of reduplication. For each of these elements in a word, there may be a correspondence within the word's translation in another language. There may be similarities in the order of the morphemes, their forms, and the morphophonological processes that take place at their boundaries. Some of these correspondences are illustrated in Figure 3a for an Am-Ti-Te translation set. The arrows indicate corresponding morphemes, with the phonetic relationships relatively clear, except in the case of the first prefix. Within the stem, the root consonants appear in boldface. Though this may not be apparent, the Am and Ti roots in this case are systematically related: the Ti consonant /ħ/ corresponds regularly to zero in Am, and the Ti consonant /s'/ corresponds often (though not always) to /t'/.

Clearly translation between words such as those in Figure 3a can benefit from recognizing the relationships indicated by the arrows in the figure. Alignment is the process of determining such relationships.

Alignment—identifying corresponding elements in source and target—is fundamental to the training of MT systems; the progress in neural MT resulting from the inclusion of attention mechanisms is just the most recent example (Bahdanau et al., 2015; Vaswani et al., 2017). Within SMT, alignment is a separate step, and because everything else depends on the quality of the alignment, a great deal of emphasis is given to alignment mechanisms (Och and Ney, 2003).

How is character alignment constrained for word translation pairs in closely related languages? With rare exceptions (see below), we can expect it to be *monotonic*. That is, for word translation pair s and t, if the character in position i in s is associated with the character in position j in t, then the character in position $i + 1$ in s is either (1) not associated with any character in t or (2) associated with a character in position k in t such that $k >= j$. This is illustrated for the Am-Ti pair in Figure 3b. Informally, monotonicity means that none of the association arrows (as in the upper diagram in Figure 3b) cross one another. Given this constraint, we can represent character alignments in the simpler format shown in the lower diagram in the figure, with aligned characters above one another.

What would it mean for the monotonicity constraint to be violated for word pairs in related languages? As far as we can tell, the only situation where this could occur is one in which one of the languages has undergone a diachronic or synchronic metathesis process, by which a pair of characters has swapped their positions. For example, imagine a word *gabla* in language 1, which translates as the word *galba* in language 2. While we are not aware of any such cases, they are certainly possible, but in this paper we restrict ourselves to the computationally simpler, and obviously much more common, monotonic cases.

Morphemes can of course consist of multiple characters, and a morpheme in one language can correspond to a morpheme of a different length in another; there are several examples in Figure 3a. However, as in Durrett and DeNero's (2013) approach to paradigm cell filling, we begin with the *alignment of single characters*, constraining the alignment to be one-to-one in both directions. As can be seen from the alignment shown in Figure 3b, these one-to-one character alignments usually do not represent morpheme correspondences, but they do represent a step towards the identification of corresponding morphemes.

Following character alignment, Durrett and DeNero merge neighboring aligned characters into candidate morphemes. This step will be the topic of the next phase of the project and a future paper.

4 Approach

Durrett and DeNero (2013) realize the alignment step in their paradigm cell filling system as an iterated edit-distance algorithm. However, the word translation problem does not lend itself to this approach because of the number of associated characters that are not identical. We applied Durrett and DeNero's algorithm to our task but were only able to achieve reasonable results by modifying it to include explicit character categories, such as consonants and vowels, affecting the probabilities of the edit operations. For example, we treated the substitution of a vowel for a vowel as more likely than the substitution of a vowel for a consonant. In general, we cannot expect to have this kind of knowledge of character categories for arbitrary orthographic systems, so it is desirable to develop a less knowledge-intensive approach to alignment. Since the early SMT work makes no assumptions at all about the categories of source and target words, we looked to this work for inspiration in designing an approach to our task that makes no assumptions about the categories of source and target characters.

4.1 IBM word alignment models

Nearly all SMT research relies on the original insights from Brown et al. (1993). Alignment is implemented through a set of parameters whose values are determined by maximizing the likelihood of a training corpus of sentence pairs. The maximization is performed using the expectation maximization (EM) algorithm (Dempster et al., 1977). In the expectation (E) step of each iteration, a set of "expected counts" of co-occurring source and target words within the corpus is calculated, weighted by the current values of the parameters. In the maximization (M) step of each iteration, new values of the parameters are calculated on the basis of the expected counts from the E-step.

4.2 Character association probabilities

In the simplest SMT alignment model (IBM Model 1), the only parameters involved are word-word translation probabilities. The goodness of an alignment of the words in a pair of sentences depends only on the probabilities that the aligned words are translations of one another, not on position in the sentences. There is the additional assumption that each word in the source sentence can be aligned with at most one word in the target sentence.

For our word translation task, we start with IBM Model 1, with words replaced by characters. That is, for each pair of source and target language characters, there is a *character association probability*, representing the probability of the source character given the target character. Each source character also has a probability of being "deleted," that is, of no association at all in the target word. However, we start with a bit more knowledge here than in the case of sentence translation. Specifically characters that are identical should have relatively high association probabilities. Thus we initialize the probabilities with a high value for identical characters (a meta-parameter that we set to 0.5) and equal low probabilities for all other character pairs. As with IBM Model 1 for sentence translation, we make the assumption that each source word character is associated with at most one target word character. In addition, we assume the inverse constraint as well: that each target word character is associated with at most one source word character. During the E-step of this simple version of our model, for each word pair in the corpus, for each combination of source and target language character, we count the ways they can be aligned (if they occur at all in the words), weighted by their current association probabilities. During the M-step we recalculate the probabilities based on the counts. The E- and M-steps are repeated until the parameters converge within some criterion.

4.3 Position and length effects

IBM Model 1 is clearly inadequate for sentence translation because a target language word is equally likely to be aligned with a source language word wherever it appears in the sentence; the order of the words is irrelevant. Succeeding models in the IBM series include additional parameters to handle position, as well as relative sentence length.

We also add additional parameters to handle position, but because our task is more constrained than the sentence translation task, we diverge here from the IBM models. Specifically, given our monotonicity constraint, we expect a character at a particular relative position within the source word to be associated with a character at a similar relative position with the target word. For example, in the pair of words in Figure 3b, even though the Am *l* and the Ti *y* are not equivalent, they should have a relatively high probability of association because of their positions within the word.

To operationalize this notion, we view an alignment as either "left-justified" or "right-justified," as proceeding from one or the other end of the word. Thus rather than relative distance within the words, we are actually comparing their absolute distances from one end of the word or the other. We include a set of parameters representing the probability of a left-justified alignment given the difference in the lengths of the two words. That is, for each of the possible differences in word length (there are about 20 of these in a typical dataset for our languages), the system is trained on a *left justification probability* (the right justification probability is just one minus this value).

During the E-step in this modified version of the model, the expected counts for each word pair depend not only on the current association probabilities for each character pair but also on the positions of the associated characters and on the left justification probabilities. That is, separate counts are calculated starting at the left and right ends of the word pairs, each weighted by the justification parameters. In addition, expected counts are also calculated for left justified and right justified alignments of each word pair, to use in updating the left justification parameters during the M-step.

Finally, the number of characters in a source word that are associated with characters in the target word depends not only on the difference in the lengths of the words but also on the number of target characters that are "deleted," that is, not associated with any source language characters. For example, consider the example in Figure 3b. The source word has nine characters, the target word ten characters. The number of deleted characters in the source word not only depends on the length difference (-1) but also on the likelihood that target characters are deleted. In particular the Ti phoneme /ħ/ is nearly always deleted. Thus we include a separate set of *target character deletion probabilities*, which are also trained during EM.

In summary, we define three sets of parameters governing alignments over word translation pairs and learn the values of the parameters that maximize the likelihood of the training set: (1) character association parameters, (2) left justification parameters, and (3) target character deletion parameters. We train the system using the EM algorithm. During the E-step, expected counts are calculated for each of the parameter types over the whole training set using the current parameter values as weights. During the M-step, new values for each of the parameters are calculated on the basis of the counts from the E-step. Following training, we align a given pair of words by searching for the alignment with the highest probability given the parameter values.

4.4 Possible data sources

Where would the sets of word translations for training the parameters come from? In the most general case, we are interested in languages with little or no available data, but for related languages there is often a community of multilinguals, especially when one of the languages is a lingua franca or an official regional or national language. Thus one source of data would be elicitation of word translations from multilinguals. In other cases, we might benefit from limited corpora. For the purposes of this paper, however, we wish to show only that the framework is viable, so we work with a set of languages for which we have the morphological and lexical resources necessary to generate the data automatically. Below, we describe the generation of the datasets in detail.

Applying the method to a language with few or no resources at all, through a combination of elicitation of translated words from multilinguals or word alignment in limited corpora, will wait for future work.

5 Experiments

The goal of the experiments reported on below is to investigate whether our EM alignment algorithm (1) results in character alignments that reflect the correspondences between affixes and root consonants

CATEGORY	Tense-Aspect	Polarity	Subordination	Subject Agr	Object Agr
PROPERTIES	perfective imperfective	affirmative negative	main clause relative clause	1s, 3sm, 2p(f)[6]	0, 3sm, 3sf
EXAMPLES					
Am: አልመከራችኋትም	perfective	negative	main clause	2p	3sf
Ti: እቕድሶ	imperfective	affirmative	main clause	1s	3sm
Te: ለኢልበሸል	imperfective	negative	relative clause	3sm	0

Table 1: Morphological properties used in generating datasets.

and (2) yields parameters that embody the known phonetic relationships between the languages.

5.1 Training and test data

We generate the data for the experiments automatically using our multilingual root lexica and the morphological generators available in HornMorpho (Gasser, 2011). Because our initial assumption is that alignment will succeed only for words with similar roots, using our root lexica for the languages, we first extract a set of verbs in the three languages with the same or similar meanings that have the same or similar roots.[4]

The result is a set of 42 roots for each of the three languages.[5] For each of the roots, we generate a set of word forms, yielding all combinations of a set of morphological properties that are common to the three languages. These properties are shown in Table 1. This results in 1488 word translation sets. We reserve 298 of these for validation and testing, yielding a training set of 1190 translation sets. Each translation set appears in an orthographic form and a phonemic form. We also isolated a separate set of 42 *dissimilar* root pairs for Am and Ti (see below for the motivation), for example, the pair Am <*č'fr*> Ti <*sʕsʕ*> 'dance'. For each of these roots, we generated word forms using the same set of morphological properties. The result is a separate dissimilar training set consisting of 1488 word translation pairs and a "mixed" training set of 2,380 pairs consisting of both similar and dissimilar pairs.

The number of distinct characters in the orthographic data sets is as follows: Am–72, Ti–76, Te–57. The phonemic data sets include separate characters for the geminated and labialized forms of consonants; for example, *k*, *kː*, k^{w}, and $k^{w}ː$ all appear in the Am data as "characters". The number of distinct characters in the phonemic data sets is as follows: Am–47, Ti–40, Te–39.

5.2 Training

For all conditions, we trained eight iterations of the EM algorithm. Following training, we aligned the test items on the basis of the learned parameters and evaluated the alignments on the basis of the constraints described below.

We first trained the Am-to-Ti and Ti-to-Te translation pairs, separately for the orthographic and phonemic training sets. Next, we trained each of these translation pairs in the opposite direction. Then, inspired by symmetrization techniques applied to alignment in SMT (Och and Ney, 2003), we aligned the test set on the basis of the union and the intersection of alignments based on the parameters learned in the two directions. In the union case, characters were considered aligned if they would have been aligned in either of the two directions. In the intersection case, characters were considered aligned if they would have been aligned in both of the two directions.

Finally, we attempted to address the artificiality of our training set by training on the "mixed" training set described above. For the mixed condition, the test set was the same one used for the similar roots condition.

[4]Phonetic "similarity" here takes into account known relationships among the languages, for example, the fact that the root consonants *h*, *ħ*, *ʔ*, and *ʕ* in Ti and Te correspond to zero in Am roots. Of course these are relationships that must be discovered by the alignment mechanism.

[5]More precisely, each "root" is a combination of a consonantal root, its root category, and one of the set of derivational categories that characterize EES verbs, for example, passive-reflexive or transitive. An example is the Ti "root" <*lʔk*>:A:ps 'be sent', consisting of root consonant sequence *lʔk* in root class A and in the passive-reflexive derivational category.

5.3 Evaluation

We annotated each phonemic and orthographic test set of 149 word pairs by hand, based on our knowledge of the morphology of the languages; see the upper portion of Figure 4. For the phonemic translation pairs, we indicate specific character-to-character alignments for root consonants, where these are present in both words, and for affixes, when there are clear correspondences between the characters. In cases where corresponding affixes do not clearly map onto one another, we constrain the characters in the two affixes to overlap. That is, each character within the shorter affix should align with some character in the longer affix. For the orthographic translation pairs, we indicate corresponding root characters (sharing a consonant but not necessarily a vowel), where these are present in both words. For affixes, annotation is as for the phonemic pairs. There are no constraints associated with affixes in one language that correspond to nothing in the other language. The upper part of Figure 4 shows an example, the Am-Ti translation pair from the previous figures, translating 'which was not washed'. Root consonant correspondences are indicated by arrows, affix correspondences by boxes. There are seven alignment constraints in the phonemic pair, five constraints in the orthographic pair.

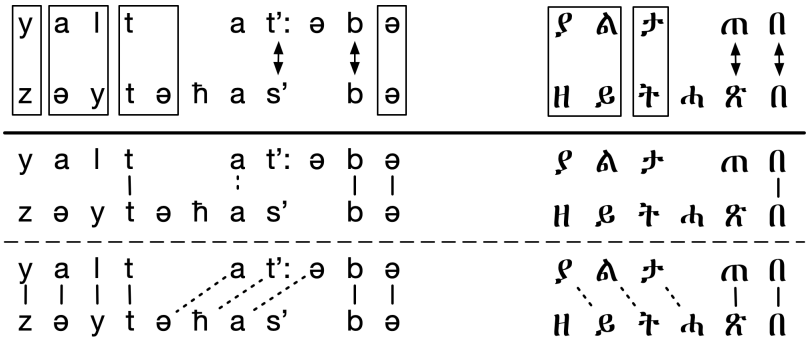


Figure 4: Alignment constraints (above), possible parameter-based alignments (below) for an Am-Ti translation pair.

We evaluated the performance of the system for both recall, the proportion of alignment constraints that were satisfied, and precision, the proportion of aligned characters that satisfied the constraints. The lower part of Figure 4 shows two alignments for the phonemic and two for the orthographic pair in the upper part of the figure. Correct alignments are indicated with solid lines, incorrect alignments with dashed lines. The upper alignments are like those we would expect early in training; only identical characters are aligned. (Recall that the character association probabilities for identical characters are initialized with much higher values than those for other character pairs.) Note that precision is high in this case: 0.75 for the phonemic pair, 1.0 for the orthographic pair. Because many correct alignments are missed, however, recall is low: 0.43 for the phonemic pair, 0.2 for the orthographic pair. The lower alignments are of a sort we would expect later in training. In both the phonemic and orthographic pairs, the system is confused because of the similarities (based on the trained parameters) of characters that should not be aligned. Recall is 0.86 for the phonemic example, 0.4 for the orthographic example. Precision is 0.67 for the phonemic example, 0.4 for the orthographic example.

5.4 Results and discussion

Results of the experiments are shown in Table 2. From these results, we may notice the following points.

First, note the relatively high baseline precision scores, especially for the orthographic data. This is not surprising, given alignments based almost entirely on character identity (see Figure 4).

Second, as we might expect, we see somewhat higher recall and precision for the more closely related languages, Ti and Te, than for Am and Ti. This suggests that the alignment mechanism might function as a measure of morphological relatedness between languages or dialects.

Third, clearly both recall and precision improve a great deal with training; this can be also be seen in the uniformly increasing F-scores. Since the system starts with no knowledge of character similarity, this is significant in and of itself. Recall improves in all four cases when alignments are formed from the union of the alignments in the two directions ("$L1 \overset{\cup}{\leftrightarrow} L2$" in the table). Somewhat surprisingly, we do not

	Orthographic						Phonemic					
	Recall		Precision		F-score		Recall		Precision		F-score	
	Base	Train	Base	Train	Base	Train	Base	Train	Base	Train	Base	Train
Am→Ti	0.405	0.906	0.865	0.959	0.552	0.932	0.581	0.925	0.650	0.813	0.614	0.865
Ti→Am	0.405	0.926	0.870	0.976	0.553	0.950	0.496	0.893	0.625	0.824	0.553	0.857
Am$\overset{\cup}{\leftrightarrow}$Ti	–	0.960	–	0.976	–	0.968	–	0.940	–	0.821	–	0.876
Am$\overset{\cap}{\leftrightarrow}$Ti	–	0.896	–	0.961	–	0.927	–	0.925	–	0.813	–	0.865
Am$\overset{\cup}{\leftrightarrow}$Ti (mixed)	–	–	–	–	–	–	–	0.894	–	0.770	–	0.827
Ti→Te	0.484	0.973	0.868	0.979	0.621	0.976	0.659	0.976	0.773	0.864	0.711	0.917
Te→Ti	0.515	0.974	0.880	0.977	0.650	0.975	0.756	0.981	0.797	0.858	0.776	0.915
Ti$\overset{\cup}{\leftrightarrow}$Te	–	0.981	–	0.980	–	0.980	–	0.989	–	0.856	–	0.918
Ti$\overset{\cap}{\leftrightarrow}$Te	–	0.973	–	0.979	–	0.976	–	0.976	–	0.864	–	0.917

Table 2: Experimental results.

see the same sort of improvement with precision, which we would expect to benefit from the intersection of the alignments ("$L1 \overset{\cap}{\leftrightarrow} L2$" in the table).

Fourth, with dissimilar roots included along with similar roots in training ("mixed" in the table, reported only for the union of the alignments and for Am and Ti translation pairs), recall is not as high as with only similar roots, but there is still significant improvement over the baseline. Of the expected correspondences between root consonants and affix segments, 0.894 are correctly aligned. Even with as many dissimilar as similar roots, the alignment algorithm is not confused.

To further investigate the value of what has been learned, we looked at the trained character association probabilities, which should reflect the morphological and phonological relationships between characters in the two languages. Focusing on the Am-Ti case, we first isolated those character pairs with character association probabilities of 0.01 or greater. We then took the intersection of the set of such characters found in the two translation directions. We performed this analysis for both the phonemic and orthographic conditions. Not surprisingly, Am characters tend to be associated with the identical Ti characters. Other intriguing relationships are also found. Because gemination operates quite differently in the two languages, ten of the consonants in the phonemic conditions are associated with their geminated variants (recall that these take the form of different characters in the data). In addition, in both the phonemic and orthographic conditions, the Am phoneme /*t*'/ is clearly associated with the Ti phoneme /*s*'/, a frequent and familiar correspondence between the languages. Similarly, the Am phonemes *k* and *k*' correspond, when ungeminated, to fricative versions of these phonemes in Ti, and again these correspondences are found, in both the phonemic and orthographic conditions. Finally, in the orthographic condition, 27 of the characters are associated with other characters in the same row in the Ge'ez character table, that is, with characters representing the same consonant followed by different vowels (see Figure 2).

6 Conclusions

The main contributions of this paper are (1) the introduction of a new task, that of translation of sets of words in morphologically complex, closely related languages and (2) the demonstration that a simple alignment algorithm, based on work within SMT, is a promising first step in approaching this task. In particular, we have shown that the parameters that are learned during EM training support alignments that reveal (1) correspondences between affixes and root consonants in pairs of EES languages and (2) relationships between orthographic or phonetic segments that we know to be valid.

Of course we have not yet shown to what extent alignment actually supports translation of morphologically complex words. To that end, our next step will be to proceed roughly as in the work of Durrett and DeNero (2013), extracting a set of transformation rules based on the alignments and then using a discriminative sequence model to learn the contexts for the rules.

References

Solomon Teferra Abate, Michael Melese Woldeyohannis, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Biniyam Ephrem, Tewodros Abebe, Wondimagegnehu Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Santa Fe, NM, USA.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels, October. Association for Computational Linguistics.

Girma A. Demeke. 2001. The Ethio-Semitic languages (re-examining the classification). *Journal of Ethiopian Studies*, 34(2):57–93.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia, June. Association for Computational Linguistics.

Ilknur Durgar El-Kahlout, Emre Bektaş, Naime Şeyma Erdem, and Hamza Kaya. 2019. Translating between morphologically rich languages: An Arabic-to-Turkish machine translation system. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 158–166.

Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash. 2020. The paradigm discovery problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7778–7790.

Michael Gasser. 2011. HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In *Proceedings of the Conference on Human Language Technology for Development*, Alexandria, Egypt.

Huimin Jin, Liwei Cai, Yihui Peng, and Xia. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy, August. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Gregory Stump. 2017. The nature and dimensions of complexity in morphology. *Annual Review of Linguistics*, 3:65–83.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.