Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 97–105
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

MultiSeg: Parallel Data and Subword Information for Learning Bilingual Embeddings in Low Resource Scenarios

Efsun Sarioglu Kayi * , Vishal Anand * , Smaranda Muresan
Columbia University
Department of Computer Science, New York, NY, USA
{ek3050, va2361, sm761}@columbia.edu

Abstract

Distributed word embeddings have become ubiquitous in natural language processing as they have been shown to improve performance in many semantic and syntactic tasks. Popular models for learning cross-lingual word embeddings do not consider the morphology of words. We propose an approach to learn bilingual embeddings using parallel data and subword information that is expressed in various forms, i.e. character n-grams, morphemes obtained by unsupervised morphological segmentation and byte pair encoding. We report results for three low resource languages (Swahili, Tagalog, and Somali) and a high resource language (German) in a simulated a low-resource scenario. Our results show that our method that leverages subword information outperforms the model without subword information, both in intrinsic and extrinsic evaluations of the learned embeddings. Specifically, analogy reasoning results show that using subwords helps capture syntactic characteristics. Semantically, word similarity results and intrinsically, word translation scores demonstrate superior performance over existing methods. Finally, qualitative analysis also shows better-quality cross-lingual embeddings particularly for morphological variants in both languages.

Keywords: low resource languages, crosslingual embeddings, byte-pair encoding, morphological segmentation

1. Introduction

Considering the internal word structure when learning monolingual word embeddings has shown to produce better quality word representations, particularly for morphologically rich languages (Luong et al., 2013; Bojanowski and others, 2017). However, the most popular approaches for learning cross-lingual embeddings have yet to use subword information directly during learning in the cross-lingual space.

One of the most widely used approaches for monolingual embeddings (fastText) (Bojanowski and others, 2017) extends the continuous skip-gram model with negative sampling (SGNS) (Mikolov et al., 2013a) to learn subword information given as character n-grams and then representing words as the sum of the n-gram vectors. SGNS has also been used to learn bilingual embeddings using parallel data, the most notable approach being BiSkip (a.k.a, BiVec) (Luong et al., 2015a). This joint model learns bilingual word representations by exploiting both the context co-occurrence information through the monolingual component and the meaning equivalent signals from the bilingual constraint given by the parallel data.

We propose a combined approach that *integrates subword information directly when learning bilingual embeddings* leveraging the two extensions of the SGNS approach. Our model extends the BiSkip model that uses parallel data by learning representations of subwords and then representing words as the sum of the subword vectors (as was done in the monolingual case for character n-grams (Bojanowski and others, 2017)). As subwords, we consider character n-grams , morphemes obtained using a state-of-the-art unsupervised morphological segmentation approach (Eskander et al., 2018) and byte pair encoding (BPE) (Sennrich et al., 2016).

We report results for learning bilingual embeddings for three low resource languages (Swahili-swa, Tagalog-tgl, and Somali-som) and a high resource language (German-deu), all of which are morphologically rich languages. For German, we simulate a low-resource learning scenario ($100K$ parallel data). Our results show that our method that leverages subword information outperforms the BiSkip approach, both in intrinsic and extrinsic evaluations of the learned embeddings (Section 3.). Specifically, analogy reasoning results show that using subwords helps capture syntactic characteristics. Qualitative and intrinsic analysis also shows better-quality cross-lingual embeddings particularly for morphological variants.

2. Methodology

Our proposed method to learn bilingual embeddings uses both parallel data and information about the internal structure of words in both languages during training. In SGNS, given a sequence of words $w_1, ..., w_T$, the objective is to maximize average log probability where c represents the context:

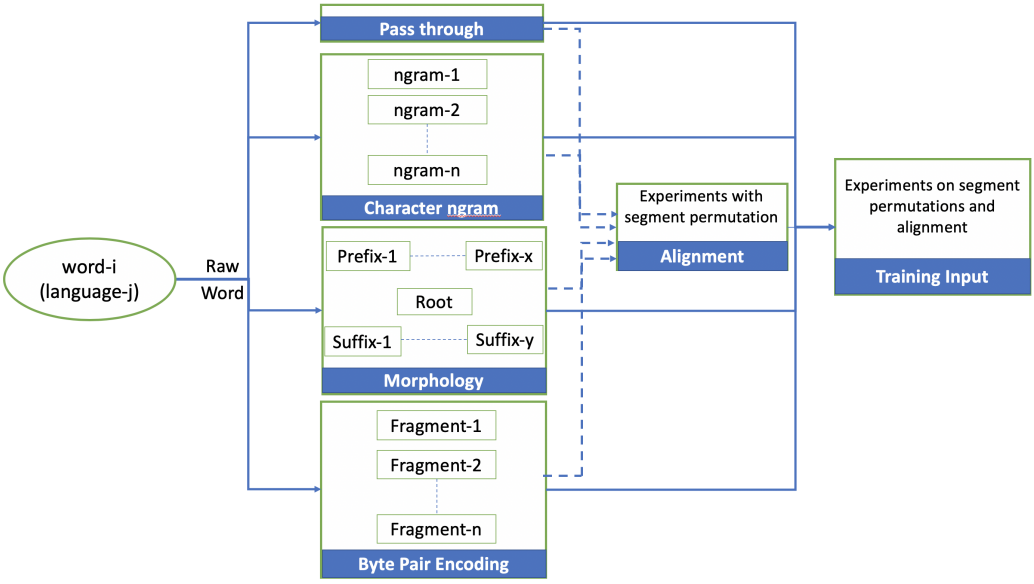
$$1/T \sum_{t=1}^{T} \sum_{c} logp(w_c|w_t), \tag{1}$$

This probability can be calculated with a softmax function as below:

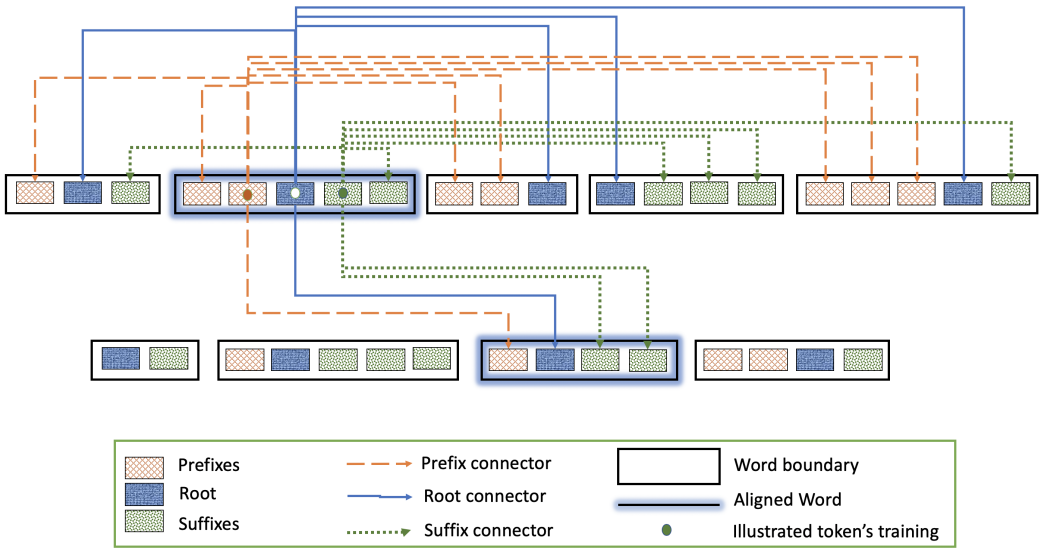
$$logp(w_c|w_t) = \frac{\sum e^{u_{w_t}^T v_{w_c}}}{\sum^{W} e^{u_{w_t}^T v_w}} \tag{2}$$

where W is the size of the vocabulary, and u_{w_t} and v_{w_c} are the corresponding word vector representations for w_c and w_t in $\mathbb{R}$. BiSkip (Luong et al., 2015b) uses sentence-level aligned data (parallel data) to learn bilingual embeddings by extending the SGNS to predict the surrounding words in each language, using SGNS for both the monolingual and cross-lingual objective. In other words, given two languages l_1 and l_2, BiSkip model trains four SGNS models

* Equal Contribution



(a) Training and Alignment Schema for word w_i in language l_j



(b) MultiSeg model illustration for Morph$_{All}$ case

Figure 1: MultiSeg Architecture

	Somali	**English**
Word	Wax aanan si fiican umaqlin ayuu ku celceliyey .	he repeated something that I could not hear well .
Stem	Wax aan si fiic maql ayuu ku celcel .	he repeat someth that I could not hear well .
Alignment	Wax:something aanan:something aanan:I si:that si:could fiican:well umaqlin:hear ayuu:not ku:NA celceliyey:repeated	

Table 1: English-Somali Alignment

jointly which predict words between the following pairs of languages:

$$l_1 \rightarrow l_1, l_2 \rightarrow l_2, l_1 \rightarrow l_2, l_2 \rightarrow l_1 \quad (3)$$

However, in this model each word is assigned a distinct vector. To take into account the morphology of words in both languages, we extend BiSkip to include subword information during learning. The approach is based on the idea introduced by Bojanowski and others (2017) for the monololingual fastText embeddings, where the SGNS is extended to learn the representation of character n-grams and then represent the word as the sum of its n-gram vectors as in Equation 4 where N is set of character n-grams and c_n is the word embedding for n-gram n.

$$w = 1/|N| \sum_{n \in N} c_n \quad (4)$$

In our approach, which we call MultiSeg, we consider subwords as character n-grams (between *3* and *6* as in fastText), or as morphemes, or as byte pair enconding (BPE)

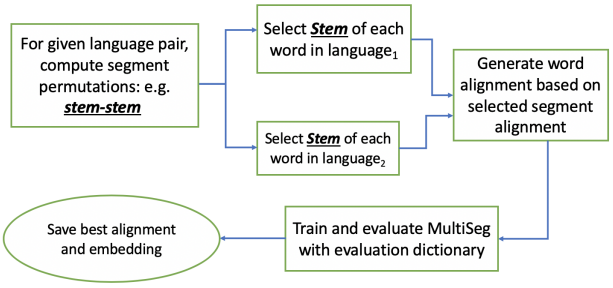


Figure 2: Alignment algorithm

that are computed by merging most frequent adjacent pairs of characters in the corpora. When considering morphemes as subwords, we either split the words into *prefix, stem* and *suffix*, or we consider all morphemes, that is the *stem* and all *affixes*. We use an unsupervised morphological segmentation approach (Eskander et al., 2018; Eskander et al., 2019) based on Adaptor Grammars that has been shown to produce state-of-the-art results for a variety of morphologically rich languages (e.g., Turkish, Arabic, and 4 Uto-Aztecan languages which are low resource and polysynthetic).

We denote our proposed method using each subword type as MultiSeg$_{CN}$ (uses char n-gram as representation during training), MultiSeg$_{M}$ (uses prefix, stem, suffix morphemes), MultiSeg$_{morph_{all}}$ (uses all morphemes), MultiSeg$_{BPE}$ (uses byte pair encondings), MultiSeg$_{All}$ (uses all subword types as representations during training). Figure 1a shows all possible segmentations for a given word in a language. Once best alignment is chosen, e.g. word-level or stem-level alignment, it is passed as input data to the training algorithm. As an example, Figure 1b shows subword structure i.e. morphological segmentation, of two parallel sentences, one in English and the other in low resource language. First sentence consists of five words and the corresponding aligned sentence consists of four words and internally, they are made up of various counts of segments i.e. one root and one or more prefixes and suffixes. For the current word in training (highlighted in the Figure 1b), corresponding aligned word in the other sentence is also highlighted and their internal alignment is shown. Similarly, within the same sentence, the current word's internal alignment with neighboring words in its context is shown. For aligning segments of the words, we consider several possibilities i.e. word and stem-based alignment and pick the best one as shown in Figure 2. Example Somali and English sentences and their stemmed output is shown in Table 1. In the case that alignment based on stem performs better than alignment based on words, word level alignment can still be constructed through stem-to-word connection.

Dataset	Parallel Sentences	Vocabulary	TD
Swahili	24,900	48,259	7,720
Tagalog	51,704	43,646	9,523
Somali	24,000	66,870	12,119
German	100,000	59,333	57,617

Table 2: Data Statistics (TD: Test Dictionary pairs)

2.1. Training of Bilingual Embeddings

This section describes the data used for training our bilingual word embeddings and our evaluation setup, including the evaluation datasets and measures.

We build bilingual embeddings for Swahili-English, Tagalog-English, Somali-English and German-English. For Swahili, Tagalog and Somali, we use parallel corpora provided by the IARPA MATERIAL program[1]. Data statistics for each language i.e. size of parallel corpora, vocabulary and dictionaries, are listed in Table 2. For German, we use the Europarl dataset (Koehn, 2005). Since the size of this parallel dataset is much larger than the others (1,908,920), we select a random subset of $100K$ parallel sentence to imitate a low-resource scenario. This is important as parallel corpora is more costly to obtain than other bilingual resources, such as dictionaries. For all the models, symmetric word alignments from parallel corpora are learned via the fast align tool (Dyer et al., 2013). For aligning segments of the words, we compute word and stem-based alignments and between the two, aligning based on stem performs better across all languages and dimensions. We train embeddings with different dimensions, $d = 40$ and $d = 300$, for 20 iterations. Our code for training MultiSeg embeddings, pre-trained cross-lingual embeddings and evaluation scripts such as word translation score and coverage will be publicly available[2].

We evaluate our approach both intrinsically and extrinsically on various monolingual and cross-lingual tasks and compare the performance to the BiSkip baseline. Recall, that BiSkip does not use any subword information when training the bilingual embeddings.

2.1.1. Intrinsic Evaluation

Word Translation Task. An important intrinsic evaluation task for learning bilingual embeddings is the word translation task a.k.a. *bilingual dictionary induction* which assesses how good bilingual embeddings are at detecting word pairs that are semantically similar across languages by checking if translationally equivalent words in different languages are nearby in the embedding space. As our evaluation dictionaries, we use bilingual dictionaries derived from Wiktionary using *Wikt2Dict* tool (Acs et al., 2013) which has polysemous entries in both directions. We generate Swahili-English, Tagalog-English. Somali-English and German-English dictionaries (the sizes are given in Table 2). We argue that these dictionaries are more reliable as evaluation dictionaries compared to Google Translate dictionaries, which are generally used only for evaluation. We calculate precision at k, where $k = 1$ and $k = 10$) (P@1, P@10) for both source-to-target and target-to-source directions and take an average of these scores as the final accuracy. We take the definition of the task from (Ammar et al., 2016). In conjunction with P@1 and P@10, we also report coverage as in (Ammar et al., 2016), given as the total number of common word pairs $(l_1, w_1), (l_2, w_2)$ that exist in both the test dictionary and the embedding, divided by size of the dictionary. The precision at 1 (P@1) score for

[1]MATERIAL is an acronym for Machine Translation for English Retrieval of Information in Any Language (Rubino, 2016)

[2]https://github.com/vishalanand/MultiSeg

Model	Dimension	German		Swahili		Tagalog		Somali	
		Coverage: 0.159		Coverage: 0.212		Coverage: 0.116		Coverage: 0.195	
		P@1	P@10	P@1	P@10	P@1	P@10	P@1	P@10
BiSkip	40	0.278	0.379	0.528	0.666	0.554	0.698	0.404	0.630
	300	0.358	0.492	0.613	0.749	0.640	0.770	0.513	0.729
MultiSeg$_{CN}$	40	0.296	0.429	**0.580**	0.728	0.624	0.774	0.440	0.708
	300	0.376	**0.566**	**0.632**	0.749	0.666	**0.828**	0.525	**0.830**
MultiSeg$_{M}$	40	**0.309**	0.438	**0.580**	**0.731**	**0.626**	**0.780**	0.451	0.704
	300	**0.382**	0.559	**0.632**	0.788	0.673	0.818	0.532	0.815
MultiSeg$_{M_{all}}$	40	0.306	0.435	0.580	**0.731**	0.625	0.778	0.449	0.701
	300	0.380	0.556	0.631	0.784	**0.674**	0.822	**0.538**	0.813
MultiSeg$_{BPE}$	40	0.294	0.421	0.575	0.719	0.595	0.750	0.449	0.682
	300	0.373	0.541	0.626	0.776	0.656	0.809	0.534	0.791
MultiSeg$_{All}$	40	0.305	**0.440**	0.570	0.726	0.611	0.778	**0.454**	**0.724**
	300	0.367	0.556	0.620	**0.798**	0.665	0.825	0.531	0.829

Table 3: Word translation scores and coverage percentages for all languages

Language	English	Deu/Swa/Tgl/Som	BiSkip	MultiSeg$_{CN}$	MultiSeg$_{M}$	MultiSeg$_{M_{all}}$	MultiSeg$_{M_{BPE}}$
German	correct	berichtigen					x
	correction	berichtigung			x	x	x
Swahili	office	afisi		x			
	officer	afisa	x	x	x	x	x
Tagalog	mine	akin			x		
	my	aking	x	x	x	x	x
Somali	approve	ansixinta				x	
	approving	ansixiyay			x		

Table 4: Qualitative Analysis: *x* show if the method correctly learned the word translation

one word pair $(l_1, w_1), (l_2, w_2)$ both of which are covered by an embedding E is 1 if $cosine(E(l_1, w_1), E(l_2, w_2)) \geq cosine(E(l_1, w_1), E(l_2, w_2^{'})) \; \forall w_2^{'} \in G^{l_2}$ here G^{l_2} is the set of words of language l_2 in the evaluation dataset, and *cosine* is the cosine similarity function. Otherwise, the score is 0. The overall score is the average score for all word pairs covered by the embedding. Precision at 10 (P@10) is computed as the fraction of the entries (w_1, w_2) in the test dictionary, for which w_2 belongs to the top-10 neighbors of the word vector of w_1.

Analogy Reasoning Task. Analogy reasoning task consists of questions of the form if *A* is to *B* then what is *C* to *D*, where *D* must be predicted. Question is assumed to be correctly answered if the closest word to the vector is exactly the same as the correct word in the question. We use the datasets for English (Mikolov et al., 2013b) which consist of 8,869 semantic and 10,675 syntactic questions. Some of the example semantic categories are *Capital City, Currency, City-in-State* and *Man-Woman* and some of the example syntactic categories are *opposite, superlative, plural nouns* and *past tense*.

Word Similarity Task. Word similarity datasets contain word pairs which are assigned similarity ratings by humans. These rankings are then compared with cosine similarity between the word vectors based on the Spearman's rank correlation coefficient to estimate how well they capture semantic relatedness. In our evaluations, we use three word similarity datasets: WordSimilarity-353 (WS353) (Finkelstein et al., 2001), Stanford Rare Word (RW) similarity dataset (Luong et al., 2013), and Stanford's Contextual Word Similarities (SCWS) dataset (Huang et al., 2012).

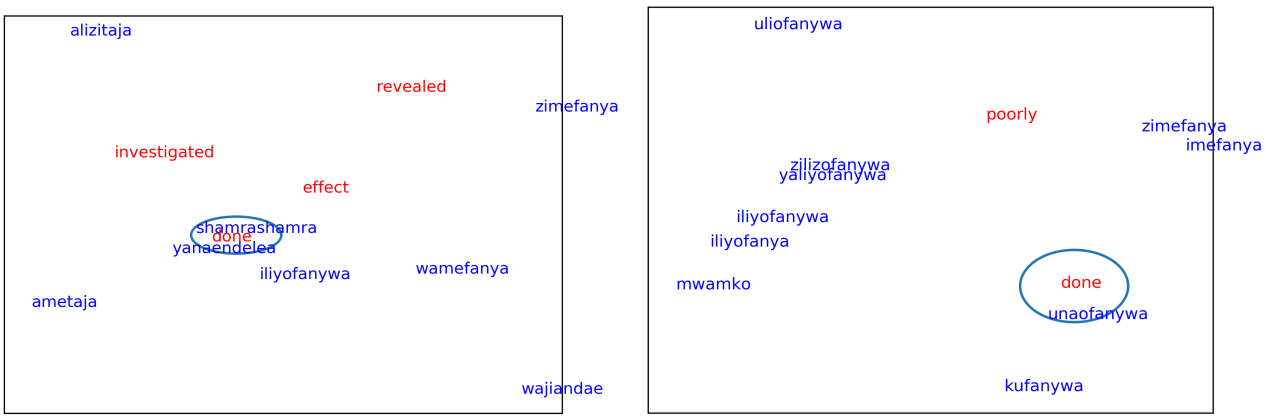
2.1.2. Extrinsic Evaluation

As extrinsic evaluation of our embeddings in a downstream semantic task, we use Cross-Language Document Classification (CLDC)[3] (Klementiev et al., 2012). In this task, a document classifier is trained using the document representations derived from the cross-lingual embeddings for language l_1, and then the trained model is tested on documents from language l_2. The classifier is trained using the averaged perceptron algorithm and the document vectors are the averaged vector of words in the document weighted by their idf values. For this task, we only have dataset for German-English, and we report results where we train on $1,000$ documents and test on $5,000$ to be consistent with the original BiSkip setup.

3. Results

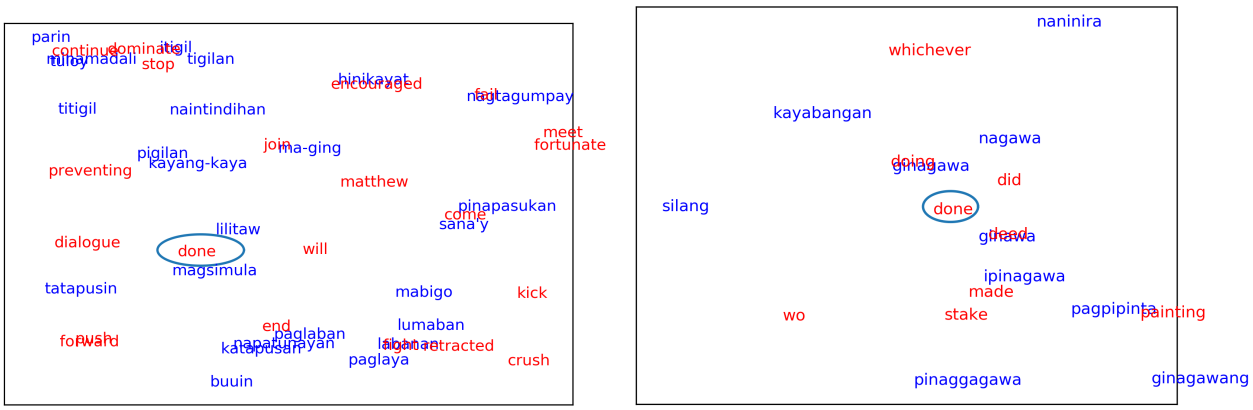
The performance on the *word translation task* for all languages is shown in Table 3, where the best scores are highlighted in red for dimension 40 and blue for dimension 300. MultiSeg methods outperform BiSkip for all languages both for P@1 and P@10. Among MultiSeg methods, across languages, morphological segmentation based models have the best scores followed by MultiSeg$_{All}$ especially for $P10$ and with 40 dimension. MultiSeg$_{CN}$ with 300 dimension also performs well across languages specifically for $P10$. Through an error analysis, we noticed that some of the performance gain for MultiSeg was due to the fact that these models were able to learn word translations of morphological variants of words. Table 4 lists some of

[3]CLDC code is provided by the authors.



(a) BiSkip (b) MultiSeg$_{CN}$

Figure 3: t-SNE visualization of English-Swahili vectors



(a) BiSkip (b) MultiSeg$_{CN}$

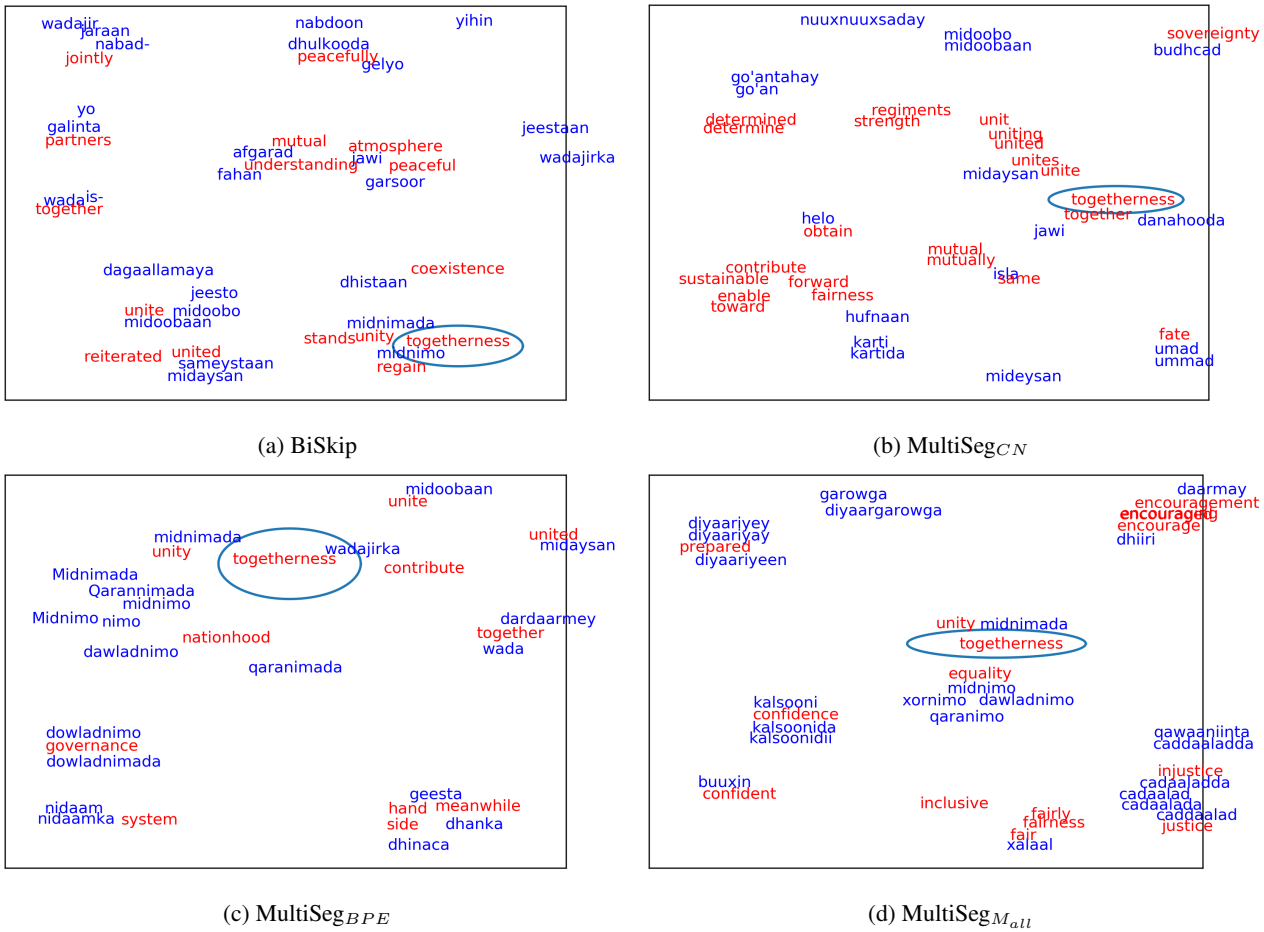
Figure 4: t-SNE visualization for English-Tagalog vectors

the examples for the words from the test bilingual dictionaries and their morphological variants and show whether or not they are predicted correctly using each technique. For all of the languages, BiSkip is only able to predict zero or one form of the word correctly, whereas MultiSeg predict various forms of the words correctly in both English and other languages.

Qualitatively, two-dimensional visualizations of cross-lingual word vectors are produced using t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008) dimensionality reduction method. Figures 3 and 4 show similar words related to the word *done* for Swahili and Tagalog respectively. It can be seen that MultiSeg$_{CN}$ learns better word representations than BiSkip by placing morphologically and semantically related words in both languages closer ($done - nagawa$, $did - ginawa$, $doing - ginagawa$). Similar graphs for Somali are provided in Figure 5 for all MultiSeg approaches. As an illustration, in Figure 5d, *qaranimo* is close to *togetherness* while the same (*nationhood*) is also shown in a coarser fashion in 5c, while other approaches could not capture this representation.

Word similarity, *analogy reasoning* and *CLDC* results for English and German are summarized in Table 5 where Spearman's rank correlation coefficients ($\rho * 100$) are reported for word similarity task and accuracy is reported for analogy reasoning task (as percentages) and for CLDC. MultiSeg approaches outperform BiSkip for all languages and for all tasks except semantic analogy. For syntactic and overall analogy reasoning scores, MultiSeg$_{All}$ performs the best which demonstrates that with better crosslingual embedding, a performance increase is seen in monolingual space, i.e. English. For CLDC task, morphological segmentation approaches, i.e. MultiSeg$_{M}$ and MultiSeg$_{M_{All}}$ perform the best. For word similarity task, overall MultiSeg$_{BPE}$ and MultiSeg$_{All}$ performs the best for English and MultiSeg$_{BPE}$ and MultiSeg$_{M_{All}}$ for German.

Word similarity and *analogy reasoning* results for English using low resource languages' cross-lingual embeddings are shown in Table 6. Again, MultiSeg approaches outperform BiSkip for all languages and for all tasks except for Somali semantic analogy and among them, MultiSeg$_{All}$ performs the best overall for all languages. A more detailed analysis of analogy reasoning task (Mikolov et al., 2013b) including breakdown of each semantic and syntactic cat-



(a) BiSkip

(b) MultiSeg$_{CN}$

(c) MultiSeg$_{BPE}$

(d) MultiSeg$_{M_{all}}$

Figure 5: t-SNE visualization for English-Somali vectors

Model	Dimension	Word Similarity				Analogy Reasoning			CLDC	
		German	English			English			eng→deu	deu→eng
		WS353	WS353	SCWS	RW	Semantic	Syntactic	All		
BiSkip	40	26.32	22.18	23.62	12.97	**3.10**	5.30	5.01	0.828	0.666
	300	25.40	22.65	21.65	8.30	**3.30**	7.74	7.16	0.839	0.667
MultiSeg$_{CN}$	40	27.60	25.77	25.91	13.64	1.20	27.56	24.11	0.814	0.662
	300	33.23	26.77	**28.68**	14.37	1.80	41.36	36.18	0.812	0.69
MultiSeg$_{M}$	40	31.10	28.48	25.61	16.44	2.80	21.64	19.18	**0.841**	0.710
	300	33.47	33.08	28.21	13.84	1.30	35.78	31.27	0.861	**0.734**
MultiSeg$_{M_{all}}$	40	31.35	30.14	26.85	16.60	2.80	22.15	19.62	0.836	**0.724**
	300	**36.00**	27.42	28.43	13.35	2.50	39.25	34.44	**0.864**	0.652
MultiSeg$_{BPE}$	40	**32.03**	**33.83**	25.51	15.11	1.70	11.28	10.03	0.812	0.720
	300	30.45	**33.64**	26.83	13.64	1.70	19.71	17.36	0.846	0.723
MultiSeg$_{All}$	40	26.97	28.59	**26.86**	**16.82**	1.20	**34.80**	**30.41**	0.822	0.631
	300	29.58	31.57	28.67	**15.52**	1.90	**48.95**	**42.79**	0.828	0.713

Table 5: German-English Monolingual and Cross-lingual Evaluation Results

egories can be seen in Figure 6 for Swahili.[4] Semantic analogy task consists of questions such as capital countries, currency, city-in-the-state and hence it does not necessarily benefit from our subword based approach. For German and Somali, BiSkip has the best performance in this category whereas for Swahili and Tagalog MultiSeg approaches perform the best. On the other hand, syntactic analogy consists of questions about base/comparative/superlative forms of adjectives, singular/plural and possessive/non-possessive forms of common nouns; and base, past and third person present tense forms of verbs. Accordingly, our representation is able to perform better for syntactical analogy questions where MultiSeg methods consistently outperform BiSkip in all of the categories. Among the MultiSeg representations, $MultiSeg_{CN}$ performs the best.

[4]We obtained similar graphs for other languages.

Language	Model	Dimension	Word Similarity			Analogy Reasoning		
			WS353	SCWS	RW	Semantic	Syntactic	All
Swahili	BiSkip	40	13.41	9.31	15.97	9.94	2.05	2.85
		300	17.25	10.05	15.64	8.01	4.28	4.66
	MultiSeg$_{CN}$	40	25.05	20.87	17.05	9.67	18.82	17.89
		300	29.43	22.06	16.62	9.39	**30.91**	**28.71**
	MultiSeg$_{M}$	40	26.05	16.92	2.97	12.43	13.79	13.65
		300	29.16	18.64	2.73	**14.64**	23.29	22.41
	MultiSeg$_{M_{all}}$	40	**27.79**	16.92	1.81	**13.54**	13.22	13.25
		300	26.19	16.48	1.99	14.09	23.67	22.69
	MultiSeg$_{M_{BPE}}$	40	26.37	19.30	2.56	11.33	7.21	7.63
		300	30.38	17.69	3.86	14.09	13.98	13.99
	MultiSeg$_{All}$	40	27.48	**21.99**	**18.24**	9.94	**20.74**	**19.64**
		300	**31.85**	**23.66**	**17.23**	11.33	29.49	27.63
Tagalog	BiSkip	40	13.17	11.49	10.37	8.64	3.11	3.67
		300	11.38	13.19	11.00	15.64	5.75	6.75
	MultiSeg$_{CN}$	40	26.18	18.64	12.80	20.78	32.54	31.35
		300	29.59	19.96	16.13	18.72	36.76	34.93
	MultiSeg$_{M}$	40	18.51	16.62	-3.11	21.60	25.23	24.86
		300	17.63	14.98	-2.80	19.14	28.66	27.70
	MultiSeg$_{M_{all}}$	40	21.08	16.24	-1.19	**26.13**	25.05	25.16
		300	20.81	17.07	-1.57	17.49	28.71	27.57
	MultiSeg$_{M_{BPE}}$	40	17.24	15.67	-1.88	24.49	14.17	15.21
		300	18.66	15.24	-1.67	**21.81**	19.60	19.82
	MultiSeg$_{All}$	40	**27.80**	**21.10**	**13.25**	21.60	**35.62**	**34.20**
		300	**28.95**	**23.21**	**14.73**	20.16	**38.31**	**36.47**
Somali	BiSkip	40	8.04	7.06	10.48	**12.82**	1.87	2.48
		300	10.92	9.86	11.96	**10.26**	2.28	2.72
	MultiSeg$_{CN}$	40	20.39	17.65	14.94	4.49	12.28	11.85
		300	**26.41**	19.02	**13.98**	2.56	22.53	21.43
	MultiSeg$_{M}$	40	16.53	10.67	-1.26	8.97	11.90	11.74
		300	15.50	11.93	0.65	5.13	24.21	23.16
	MultiSeg$_{M_{all}}$	40	15.83	9.44	-1.55	5.77	13.21	12.80
		300	16.44	12.86	-0.72	3.21	27.66	26.31
	MultiSeg$_{M_{BPE}}$	40	**21.63**	10.12	1.96	3.21	2.47	2.51
		300	19.62	11.28	0.76	4.49	4.90	4.88
	MultiSeg$_{All}$	40	20.77	**20.03**	**15.11**	7.05	**16.77**	**16.23**
		300	25.35	**19.86**	13.88	1.92	**29.19**	**27.69**

Table 6: Monolingual English Evaluation of Low Resource Languages

4. Related Work

4.1. Monolingual Morphological Embeddings

There are several ways of incorporating morphological information into word embeddings. One approach adapted by fastText embeddings (Bojanowski and others, 2017) is to use character n-grams. In addition to whole words, several sizes of n-grams, i.e. three to six, are used during training of the skip-gram model. This approach is language-agnostic and can be adapted to new languages easily. Another approach is to have morphological segmentation as a preprocessing step before training the embeddings (Luong et al., 2013). Other techniques predict both the word and its morphological tag (Cotterell and Schütze, 2015) however, all these approaches are monolingual and work on one language at a time.

The most closely related work to ours is (Chaudhary et al., 2018) which uses the fastText (Bojanowski and others, 2017) approach to include morphological information when learning cross-lingual embeddings by combining the high-resource and low resource corpora and training using the skip-gram objective. Their evaluation is limited to named-entity-recognition and machine translation and requires detailed linguistically tagged words on a large monolingual corpus for related languages. Our approach incorporates supervision through small amount of parallel corpora while training on subwords for any two languages including unrelated ones.

4.2. Bilingual Embeddings

Bilingual word embeddings create shared semantic spaces in multi-lingual contexts and can be trained using different types of bilingual resources. Techniques such as BiSkip (Luong et al., 2015b) use sentence aligned parallel corpora, whereas BiCVM (Vulić and Moens, 2015) use document aligned comparable corpora. There are also techniques that map pre-trained monolingual embeddings into shared space via bilingual dictionaries (Lample et al., 2018b; Artetxe et al., 2018). Finally, there are semi-supervised and unsupervised methods that require little to none bilingual supervision (Lample et al., 2018a; Artetxe and others, 2018). Among these techniques, we adapted BiSkip to learn embeddings jointly. This eliminates the need for having pre-

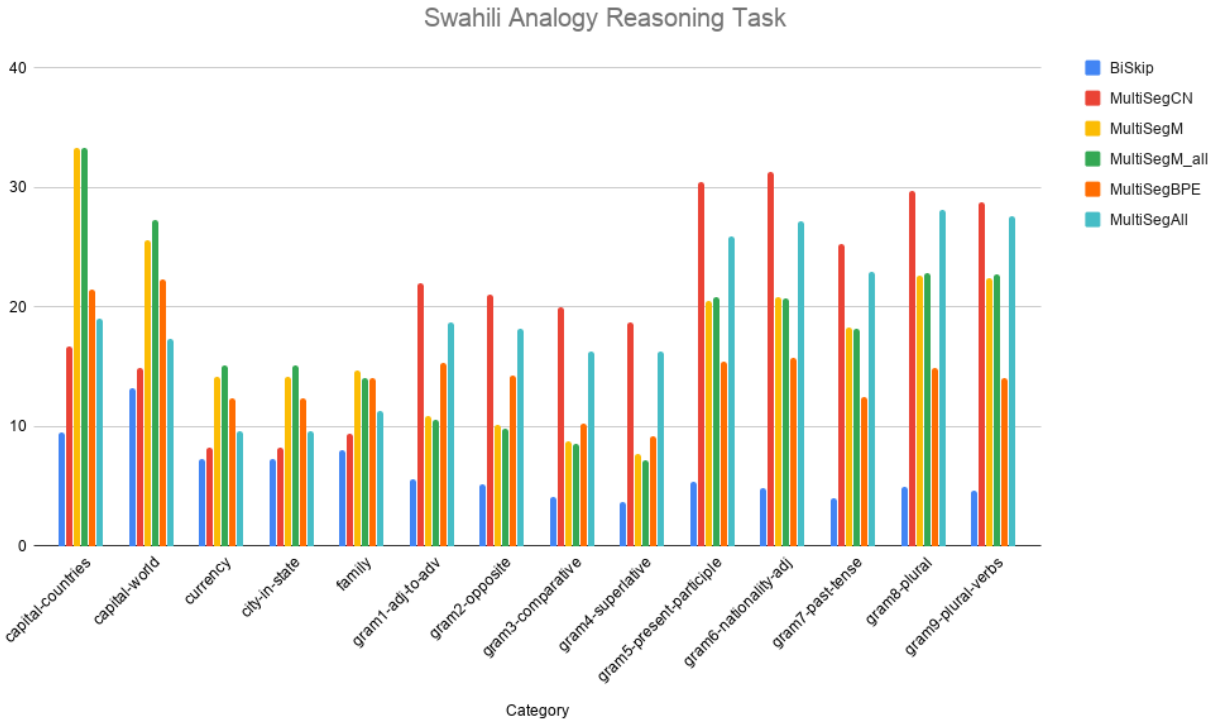


Figure 6: Swahili Analogy Reasoning Task Semantic and Syntactic Categories

trained monolingual embeddings and it has been shown to have better accuracy than comparable corpora based approaches (Upadhyay et al., 2016). In addition, our intrinsic evaluations of semi-supervised and unsupervised embeddings did not perform well.

Recently, pre-trained contextual embeddings have been extended to other languages, e.g. XLM (Lample and Conneau, 2019), cross-lingual ELMo (Schuster et al., 2019) and multilingual BERT (Devlin et al., 2019) shown to have promising results on a variety of tasks. However, they are not as amenable in low resource scenarios where they tend to overfit. They are also not good at fine-grained linguistic tasks (Liu et al., 2019) and geared toward sentence level tasks. In addition, if a pretrained model is not available, it requires lots of computing power and data to be trained from scratch. For instance, XLM model uses $200K$ for low resource and 18 million for German. For parallel data, they use $165K$ for Swahili and 9 million for German.

5. Conclusions and Future Work

We present a new cross-lingual embedding training method for low resource languages, MultiSeg, that incorporates subword information (given as character n-grams, morphemes, or BPEs) during training from parallel corpora. The morphemes are obtained from a state-of-the-art unsupervised morphological segmentation approach. We show that it consistently performs better than the BiSkip baseline, including on word similarity, syntactical analogy and word translation tasks across all languages. Extrinsically, cross-lingual document classification scores also outperform BiSkip. Finally, qualitative results show that our approach is able to learn better word-representations especially for morphologically related words in both source and target language. We plan to extend our technique to train on more than two languages from the same language family.

Acknowledgments

This research is based upon work supported by the Intelligence Advanced Research Projects Activity (IARPA) MATERIAL program, via contract FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of IARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

6. Bibliographical References

Acs, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58.

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively multilingual word embeddings. *ArXiv*, abs/1602.01925.

Artetxe, M. et al. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, pages 789–798.

Artetxe, M., Labaka, G., and Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations.

Bojanowski, P. et al. (2017). Enriching word vectors with subword information. *TACL*, 5:135–146.

Chaudhary, A., Zhou, C., Levin, L., Neubig, G., Mortensen, D. R., and Carbonell, J. (2018). Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295.

Cotterell, R. and Schütze, H. (2015). Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Eskander, R., Rambow, O., and Muresan, S. (2018). Automatically tailoring unsupervised morphological segmentation to the language. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 78–83.

Eskander, R., Klavans, J., and Muresan, S. (2019). Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, pages 406–414.

Huang, E., Socher, R., Manning, C., and Ng, A. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 873–882.

Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018a). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and JÃ©gou, H. (2018b). Word translation without parallel data. In *International Conference on Learning Representations*.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1073–1094.

Luong, T., Socher, R., and Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Luong, T., Pham, H., and Manning, C. D. (2015a). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Luong, T., Pham, H., and Manning, C. D. (2015b). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Rubino, C. (2016). Iarpa material program.

Schuster, T., Ram, O., Barzilay, R., and Globerson, A. (2019). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1599–1613.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1661–1670.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Vulić, I. and Moens, M.-F. (2015). Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 719–725.