Deep Neural Networks at the Service of Multilingual Parallel Sentence Extraction

Ahmad Aghaebrahimian
University of Innsbruck
Department of Translation Studies
Herzog-Siegmund-Ufer 15, A-6020 Innsbruck, Austria
Ahmad.Aghaebrahimian@uibk.ac.at

Abstract

Wikipedia provides an invaluable source of parallel multilingual data, which are in high demand for various sorts of linguistic inquiry, including both theoretical and practical studies. We introduce a novel end-to-end neural model for large-scale parallel data harvesting from Wikipedia. Our model is language-independent, robust, and highly scalable. We use our system for collecting parallel German-English, French-English and Persian-English sentences. Human evaluations at the end show the strong performance of this model in collecting high-quality parallel data. We also propose a statistical framework which extends the results of our human evaluation to other language pairs. Our model also obtained a state-of-the-art result on the German-English dataset of BUCC 2017 shared task on parallel sentence extraction from comparable corpora.

Title and Abstract in German

Tiefe Neuronale Netze im Dienste der Extraktion mehrsprachiger paralleler Satzpaare

Wikipedia ist eine überaus wertvolle Quelle von mehrsprachigen Paralleldaten, die für eine Vielzahl von theoretischen und praktischen sprachbezogenen Fragestellungen benötigt werden. Wir stellen ein neuartiges neuronales End-to-End-System für das massenhafte Sammeln von Paralleldaten aus der Wikipedia vor. Das System ist sprachenpaarunabhängig, robust und weist eine hohe Skalierbarkeit auf. Wir nutzen es zur Extraktion von parallelen Satzpaaren in den Sprachenpaaren Deutsch-Englisch, Französisch-Englisch und Persisch-Englisch. Die hohe Genauigkeit unseres Systems wird durch manuelle Evaluation bestätigt. Darüber hinaus stellen wir einen statistischen Ansatz vor, mit dessen Hilfe menschliche Qualitätsurteile auf weitere Sprachenpaare übertragen werden können. Unser System erzielt State-of-the-Art-Ergebnisse gemessen am deutsch-englischen Datensatz der BUCC 2017 Shared Task zur Extraktion von parallelen Satzpaaren aus Vergleichskorpora.

1 Introduction

Parallel texts are an important resource in Natural Language Processing (NLP) applications and tasks. From Statistical and Neural Machine Translation (SMT, NMT) (Brown et al., 1990; Och and Ney, 2002; Kalchbrenner and Blunsom, 2013; Cho et al., 2014), to automatic lexical acquisition (Gale and Church, 1993; Melamed, 1997), cross-lingual Information Retrieval (Davis and Dunning, 1995; Oard, 1997) and annotation projection (Yarowsky et al., 2001; Diab and Resnik, 2002) all are dependent on parallel data.

Generating parallel corpora from scratch is a highly time consuming and expensive task. Therefore, many studies focus on extracting parallel texts from comparable corpora (Munteanu and Marcu, 2005) such as Wikipedia.

Wikipedia is a useful source of parallel sentences since humans already annotated its comparable documents. In Wikipedia, one can find both parallel, and comparable articles[1]. The reason is that some

[1]In a parallel bilingual corpus like Europarl (Koehn, 2005), all sentences in source and target languages are parallel (i.e., translations of each other) while in two comparable corpora the same topic may be described with entirely different sentences.

Proceedings of the 27th International Conference on Computational Linguistics, pages 1372–1383
Santa Fe, New Mexico, USA, August 20-26, 2018.

authors prefer to explain an issue in their own words, while others use existing articles to translate them into their languages. Still, there are others who may combine the two approaches discussed above by borrowing some sentences from a source article and add some other contents by themselves. Generally, even in comparable articles with partial translations, there is a high chance of finding parallel sentences, since both articles are talking about the same topic.

A simple way of retrieving parallel sentences from comparable articles is to align the sentences in source and target pages together using a sentence alignment algorithm (Gale and Church, 1993; Fung and Church, 1994; Wu, 1994; Moore, 2002). However, these aligners are designed to align parallel corpora in which the source and target sentences are in the same order (i.e., no cross-alignment) or in proximity to each other and in which each sentence has only one matching sentence (i.e., no many-to-many alignment). These assumptions are largely violated in comparable corpora.

To address these issues, we designed a novel neural model which estimates a global probability distribution given each pair of sentences in comparable documents. We train an alignment model using German-English parallel data from the Europarl corpus (Koehn, 2005) and use the trained model to extract parallel sentences for the German-English, French-English, and Persian-English language pairs from Wikipedia.

Our model achieved statistically significant improvement over two baselines on Europarl test data (please see Section 5). Since we assumed we had no access to any gold parallel data in Wikipedia, we used human evaluation to determine the performance of our model using a two-tier evaluation design. Still, to make our results more comparable, we applied our model on the German-English dataset of BUCC 2017 second shared task on parallel sentence extraction from comparable corpora (Zweigenbaum et al., 2017) and we obtained a state-of-the-art result on it too.

In this work, we intend to model parallel sentences as accurate as possible. Hence we consider two sentences as parallel only if they have the same semantic content (i.e., convey the same message) and do not have any more or less content that is mentioned in one and missing in another (e.g., an extra or missing prepositional phrase). The sentences that do not satisfy this requirement are considered partial parallel sentences. The accuracy with which each of these partial parallel sentences represents the meaning of their source sentence is expressed in terms of a Normal distribution which is discussed in Section 6.

Moreover, we treat parallel sentences asymmetrically since professional translators often translate from their second language to their native language. Even when the translator's competency level in the source and target languages are the same, the target language could be influenced by the source language. Therefore, we trained all our models on language pairs whose target language is always English and recruited native English speakers for our evaluation tasks.

2 Related Work

Parallel data are considered an asset both in theoretical (e.g., contrastive corpus linguistics, translation studies, language use, and change) and applied (Machine Translation (MT), word sense disambiguation, bilingual lexicography) computational linguistics. There is a wealth of studies on the extraction of parallel data from the Internet in general and Wikipedia in particular. (Adafre and de Rijke, 2006) were among the first researchers who used Wikipedia for parallel data extraction. They generated a pack of source and target documents as the Cartesian product between 30 Wikipedia pages and utilized an MT system to translate target pages into English. Then, they used a similarity measure based on word overlap between the source and target sentences. In another approach, they used matching hyperlinks in Wikipedia pages to identify similar sentences.

To decrease the search space in (Adafre and de Rijke, 2006) work which is evidently too big for large-scale data extraction projects, (Mohammadi and GhasemAghaee, 2010) integrated the idea of length-based sentence alignment in (Gale and Church, 1993)[2] as a heuristic to decrease the complexity of the algorithm.

For larger-scale studies, (Barbosa et al., 2012) used bilingual dictionaries and online translation services (e.g., Google Translate or Microsoft Bing) and (Zhang et al., 2006) proposed the use of aligners

[2]Long source sentences are usually translated into long target ones and short source sentence to short target ones.

Language Pairs	Comparable articles
French-English	1,491,578
German-English	1,247,102
Persian-English	866,408

Table 1: The number of interlanguage links in Wikipedia for selected language pairs. English has interlanguage links with more than 300 other languages.

for content similarity estimation in candidate parallel web pages. (Štromajerová et al., 2016) enhanced Zhang et al.'s system by using Wikipedia's translation templates to locate comparable Czech-English parallel pages and subsequently by using the Hunalign tool (Varga et al., 2005) to extract parallel sentences.

In large-scale data extraction projects, checking all possible sentences for all pages in two languages is neither feasible nor necessary when document-level alignments are already available. (Smith et al., 2010) and (Ştefănescu and Ion, 2013) did their studies on document-aligned articles of Wikipedia. Smith et al. used a feature-based model on aligned documents. Similarly, Ştefănescu and Ion used cross-lingual Wikipedia links embedded within the documents and a trainable model to generate similarity scores for parallel sentence identification.

Classifiers are used for parallel sentence detection as well. (Chu et al., 2014) studied the use of classifiers for parallel sentence identification. They proposed a filtering scheme for Chinese-Japanese language pairs and used a binary classifier on the pruned sentences for parallel sentence classification.

In our work, we let a deep neural architecture learn the most relevant features on its own. We use Interwiki links available in Wikipedia to locate comparable pages. Using an end-to-end deep neural model we extract the most likely parallel sentences given two comparable pages by projecting the sentences into n-dimensional space. In this way, the model learns the most relevant features on its own without knowing much about the source and target languages.

3 Dataset

Wikipedia is an online encyclopedia of human knowledge. As of December 2017, it hosted over 14 million articles in more than 300 languages. While English as the biggest Wikipedia contains more than 3 million articles and 14K active users, there are 28 languages with more than 100K and 60 languages with more than 10K articles. Wikipedia is a crowd-sourced resource of information authored and translated collaboratively on a non-profit basis. Wikipedia provides a collection of similar pages in different languages by linking them together with interlanguage links. These links appear either in a sidebar on the left side or in the text of a page as in-line links. Table 1 represents the number of available English links for German, French, and Persian languages.

Our model is trained on parallel sentences from Europarl and is used to extract parallel sentences from German-English, French-English, and Persian-English comparable pages in Wikipedia. To train our model, we compiled a dataset of first 200K parallel German-English sentences from Europarl. The German sentences were translated using Google translate service and were used as pointers to original target sentences. Online translation in this phase is not a crucial step since a simple word replacement utilizing a dictionary can do the job.

We preprocessed all textual data including source and target sentences by eliminating all non-alphanumeric characters. Numeric characters were changed to 9 to retain the numeric semantic value of numeric tokens. We used 90%, 5%, and 5% of the sentences in our dataset to compile training, validation and test sets respectively in the following way.

Similar to (dos Santos et al., 2014), our model is trained by contrasting positive and negative parallel sentences. Therefore, to compile our training set, given each source sentence, we generated a positive parallel pair by taking its correct target sentence and a negative parallel pair by taking a randomly chosen sentence from target sentences. We make sure that the randomly selected sentence is not the same as the correct target sentence. To compile the validation and test sets, given each source sentence, we take a context of ten surrounding sentences including the correct target sentence. We used these sets to train,

validate and test the alignment system.

When the alignment system is trained and optimized, it is ready to extract parallel sentences from Wikipedia. For parallel sentence extraction, in the first step, we need to find comparable data in Wikipedia. Wikipedia connects comparable pages using interlanguage links. These links are available as an SQL database containing pointers to comparable pages. In Table 1 a few language pairs with their available comparable pages are reported. Using this database, we extract the page contents and use some simple preprocessing tasks to extract sentences from the pages by removing images, tables, graphs, formula, etc. In the end, we package source and target sentences of a comparable pair in one packet.

For each packet, our trained model estimates a probability distribution over all target sentences given each source sentence. In Wikipedia, we do not have access to standard gold data (i.e., we do not know which source-target sentence combination is parallel) hence, we use human evaluation (see Section 6) to estimate the system performance. We report the results of these experiments in Sections 5 and 6

4 Model Architecture

To compare source and target sentences in the mathematical sense, in the first step, we need to project them into n-dimensional space. To do this, we made use of Recurrent Neural Network (RNN) (Elman, 1990) architectures to encode textual strings into vector representations. Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014) are two widely studied variants of RNNs. In our study, we used LSTM cells since they showed better and more stable performance in our experiments.

To provide LSTM layers with their inputs, in the first layer, we used a lookup table to cast words into word embeddings (Equation 1). $\mathbf{W}$ in Equations 1 is the one-hot representation of the words in sentences whose production with pre-trained embedding matrix $\mathbf{E}$ generates word vector $\mathbf{W}_{i,t}$ for the words in the i$^{\text{th}}$ sample sentence each in time step t.

In the next layer, a forward RNN layer accepts word vectors and generates a sequence of vectors for each time step. A similar RNN does the same job but in opposite order to generate backward RNN vectors. We did max pooling ($\mathbf{MP}$) (Equations 3 and 5) over RNN vectors to get the most relevant features and then concatenated them in Equation 6. The final result of this process, $\mathbf{S}$ is a forward and backward vector representation of textual strings. We used this architecture to encode our source sentences.

$$\mathbf{W}_{i,t} = \mathbf{E}^{\intercal}\mathbf{W}_k \tag{1}$$

$$\overrightarrow{\mathbf{S}}_{i,t} = RNN(\overrightarrow{\mathbf{S}}_{i,t-1}, \mathbf{W}_{i,t}) \tag{2}$$

$$\overrightarrow{\mathbf{S}}_{i} = \mathbf{MP}(\overrightarrow{\mathbf{S}}_{i,t}) \tag{3}$$

$$\overleftarrow{\mathbf{S}}_{i,t} = RNN(\overleftarrow{\mathbf{S}}_{i,t+1}, \mathbf{W}_{i,t}) \tag{4}$$

$$\overleftarrow{\mathbf{S}}_{i} = \mathbf{MP}(\overleftarrow{\mathbf{S}}_{i,t}) \tag{5}$$

$$\mathbf{S}_i = [\overrightarrow{\mathbf{S}}_{i}; \overleftarrow{\mathbf{S}}_{i}] \tag{6}$$

Target sentences are encoded like source sentences with an additional attention layer, which helps the encoder to recognize the most relevant features by emphasizing on critical points of the target sentence given each source sentence. Likewise, the target language encoder receives the initial embeddings from a lookup table over a pre-trained embedding matrix. The forward RNN in the next layer transforms the embeddings into a sequence of vectors which finally are fed into an attention layer with attention on source sentence vectors (Equation 7). The resulted vector then is max pooled to be eliminated from non-relevant and useless features. The same process is done for the backward RNN, and the resulting vectors are concatenated.

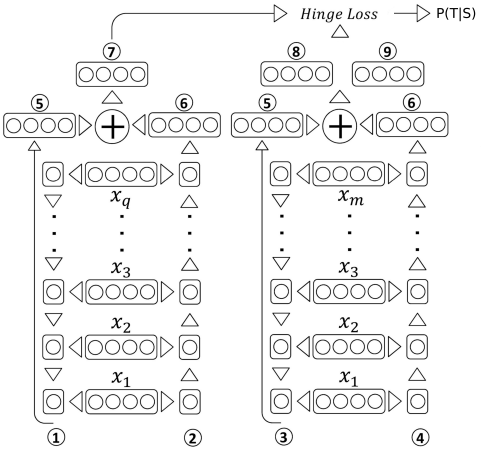


Figure 1: Abstract model architecture. Numbered components in the figure are 1-Forward LSTM, 2-Backward LSTM, 3-Forward Attentive LSTM, 4-Backward Attentive LSTM, 5-Backward max-pooling, 6-Forward max-pooling, 7-Question vector, 8-Positive sample vector, 9-Negative sample vector.

$$\begin{aligned}
\mathbf{W}_{i,t} &= \mathbf{E}^{\intercal}\mathbf{W}_k \\
\overrightarrow{\mathbf{T}}_{i,t} &= RNN(\overrightarrow{\mathbf{T}}_{i,t-1}, \mathbf{W}_{i,t}) \\
\overrightarrow{\mathbf{T}}_{i,t} &= ATT(\overrightarrow{\mathbf{T}}_{i,t}, \mathbf{S}_i) \qquad (7) \\
\overrightarrow{\mathbf{T}}_{i} &= \mathbf{MP}(\overrightarrow{\mathbf{T}}_{i,t}) \\
\overleftarrow{\mathbf{T}}_{i,t} &= RNN(\overleftarrow{\mathbf{T}}_{i,t+1}, \mathbf{W}_{i,t}) \\
\overrightarrow{\mathbf{T}}_{i,t} &= ATT(\overleftarrow{\mathbf{T}}_{i,t}, \mathbf{S}_i) \qquad (8) \\
\overleftarrow{\mathbf{T}}_{i} &= \mathbf{MP}(\overleftarrow{\mathbf{T}}_{i,t}) \\
\mathbf{T}_i &= [\overrightarrow{\mathbf{T}}_i; \overleftarrow{\mathbf{T}}_i]
\end{aligned}$$

All target sentences including positive and negative sentences are encoded using this procedure. In the end, source sentence vectors, positive target vectors, and negative target vectors are ready to be used as the inputs of a Hinge objective function. Using this function, we try to maximize the similarity in parallel sentences while minimizing the similarity in non-parallel ones. Therefore, the next step is to measure the similarity between sentences in a parallel set.

The similarity between two vectors can be estimated via different approaches such as Jaccard, Cosine, Polynomial or Manhattan, etc. However, we got better results using the Geometric mean of Euclidean and Sigmoid Dot product (GESD) proposed by (Feng et al., 2015). GESD (Equation 9) combines the angular and forward-line semantic distance between two vectors.

$$SIM(V1, V2) = \frac{1}{1+exp(-(V1 \cdot V2))} * \frac{1}{1+||V1-V2||} \qquad (9)$$

To distinguish parallel sentences from non-parallel ones, we need to train their vectors in a way that increases the similarity for parallel and decreases it for non-parallel sentences. Hinge objective function does the trick for us (Equation 10).

$$\ell = \sum_i \max(0, m + SIM(\mathbf{S}_i, \mathbf{T}_i^-) - SIM(\mathbf{S}_i, \mathbf{T}_i^+)) \qquad (10)$$

After training, the model estimates a probability for each pair of source and target sentences (i.e., $p(target|source)$). In the next section, we describe how we use these probabilities to distinguish parallel sentences from non-parallel ones.

5 Results

We trained our model on the dataset compiled from Europarl parallel data. We used 128-dimensional LSTMs for all RNNs in our model, and for the embedding layer, we used GloVe word vectors (Pennington et al., 2014). To prevent the model from over-fitting, we set the drop-out rate to 0.5 for the last layer

in each module. We used an attention model similar to the one proposed by (dos Santos et al., 2014). The model was trained on two GPUs and converged after around 3 hours of training. We used random assignment and the Bleualign tool (Sennrich and Volk, 2011) as two baselines. The results of the model on the test and validation sets are reported in Table 2. These results are the accuracy of the system for sentence alignment on the German-English dataset. Since our model is trained with source sentences translated into English, the same encoder can be used for any other source languages as long as we use the same mechanism for translation.

Up to this point, we trained our aligner. In the next step, we use this alignment system to extract parallel sentences. To estimate the performance of the model on parallel sentence extraction from Wikipedia, we designed a two-tier human evaluation, which is described in the next section.

Europarl German-English	**Baseline**	**Bleualign**	**This work**
Validation set	11.94 %	92.45 %	**96.83 %**
Test set	11.34 %	93.05 %	**97.24 %**

Table 2: Comparative evaluation of German-English alignment system. The baseline is random target sentence assignment.

6 Human Evaluation

As described in Section 3, our comparable data extraction yields more than one million comparable packets for the German and French and around 800K packets for the Persian languages (Table 1). Each comparable packet consists of two files, one of which contains source text lines and the other includes target text lines.

For each comparable packet, the model as described in Section 4 estimates a probability distribution over all target sentences given each source sentence. Each source sentence and the target sentence with the highest probability forms a probable parallel pair.

Not having access to standard gold data in Wikipedia, we do not know which of these pairs are, in fact, parallel, irrespective of the probability estimated by the model.

To estimate a threshold for the probabilities and to validate it, we designed a two-step human evaluation procedure. In the first step, we establish a threshold for the model-generated probabilities above which a sentence pair could be considered a parallel pair. Irrespective of the size of each comparable packet, this probability is a global metric of how much two sentences are semantically correlated. So we only need to establish a threshold and validate it using statistical inference. After that, we can accept source and target sentences with probabilities above the threshold as correct parallel sentences.

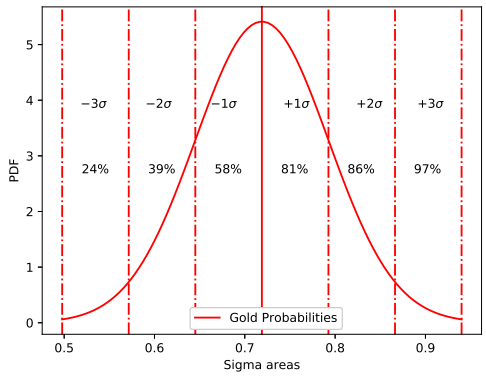
In the second step, we validate our thresholds by asking our evaluators to decide which of the extracted pairs are parallel. For this purpose, we randomly extract some sentence pairs under the curve of a Normal distribution parametrized by μ and σ obtained from the last step and ask some evaluators to decide which pairs are parallel and which ones are not.

Based on the information obtained from the evaluators, we try to reject the null hypothesis of our study (i.e., sentence pairs with a probability above the thresholds are not parallel) and to analyze the errors qualitatively.

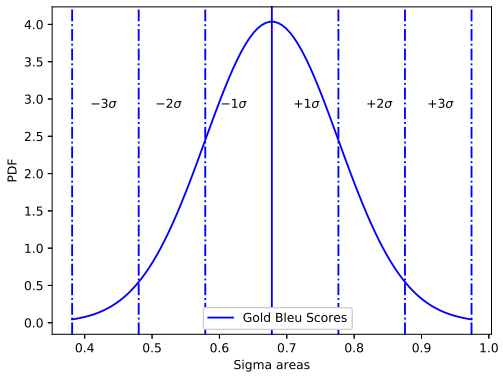
6.1 Establishing a Threshold

To establish a threshold for the model, we randomly extracted two hundred German sentences and asked two native English and fluent German translators to translate them into English. We returned these translations and their German source sentences into their original containing packets to provide them with their original contexts. Knowing that these are correct parallel sentences, we used our model to estimate their probabilities. Then, using these probabilities, we computed the μ and σ of the samples based on which we estimated a Normal distribution on the probability range of parallel sentences (Figure 2a).

Moreover, using Google translate service, we converted these sentences into English and computed their BLEU scores (Papineni et al., 1993) by using our gold translations produced by our translators as



(a) Normal Distribution on System Probabilities



(b) Normal Distribution on BLEU Scores

Figure 2: (a): Normal distribution parametrized by $\mu = 0.71$ and $\sigma = 0.07$ obtained from gold parallel sentences translated by human translators and scored by the model. The percentage in each σ region is the ratio of the sentence pairs evaluated as parallel in the second task.
(b): Normal distribution ($\mu = 0.67, \sigma = 0.09$) over the BLEU scores of English translations of 200 true parallel target sentences done by Google translate service. BLEU scores are transformed to percent.

the reference. We calculated the average of these two scores and used them for estimating a Normal distribution by computing their μ and σ (Figure 2b).

In any further parallel sentence extraction, to estimate the performance of the model without human evaluation, we only need to translate target sentences into English, to compute their BLEU score using source sentences as a reference and to compare them with the BLEU score curve obtained above. In the next section, we use this method to estimate the quality of extracted parallel sentences for French-English and Persian-English language pairs.

	-3σ	-2σ	-1σ	1σ	2σ	3σ
Observed Agreement	95 %	76 %	71 %	76 %	85 %	98 %
Cohen's kappa	85 %	69 %	45 %	50 %	81 %	89 %
Prediction Reliability	Strong	Average	Weak	Average	Strong	very Strong
Model probability Range	50%-57%	57%-64%	64%-71%	71%-78%	78%-85%	85%-92%
Bleu Score Range	40%-49%	49%-58%	58%-67%	67%-76% %	76%-85%	85%-94%
Collected parallel Sentences	24 %	39 %	58 %	81 %	86 %	97 %

Table 3: Human evaluation results. The Observed Agreement is the ratio of sentence pairs which were evaluated the same, either as true or false parallel pairs by both evaluators. Cohen's Kappa integrates chance agreement into the Observed Agreement. As mentioned in Prediction Reliability row, σ regions with Cohen's Kappa less than 0.8 are not reliably evaluated. Model Probability Ranges are the probabilities estimated by the model for 200 correct parallel sentences compile by our evaluators, and the BLEU Score Range is the BLEU scores of these sentences when translated using Google Translate. The ranges for last two rows are computed using a Normal distribution on their data. Collected Parallel Sentences are the ratio of sentence pairs which are evaluated as parallel in each σ region.

6.2 Validating the Thresholds

In Figure 2a the Normal distribution and the percentage of parallel data in each σ regions are illustrated. To validate these thresholds, we randomly and evenly sampled 1000 sentences from all σ regions and asked our evaluators to determine whether they are parallel or not. We statistically analyzed these data to assess the validity of the outcomes.

Given that our data in this task are nominal and have no order and that it is necessary to take into account the probability of chance in evaluation, we used Cohen's kappa to estimate the inter-rater reliability. We calculated the observed agreement probability for each σ region as well. The data analysis is done by considering $p > 0.95$ and 1000 data samples. The results are reported in Table 3. As is shown in this table, in $2\sigma^+$ and $3\sigma^+$ regions we have 81% and 89% inter-rater reliability accordingly, which indicates strong predictability in these regions.

The observed agreement and collected parallel sentences in these regions are quite high. In $3\sigma^+$ and $2\sigma^+$ regions, we managed to reject the null hypothesis (i.e., that all sentences in these regions are non-parallel), hence confirming that with 95% confidence, the established thresholds in these regions are statistically sufficient enough to decide whether a pair is parallel or not.

Since the thresholds are computed using the same scheme for both languages, the obtained results are valid for all language pairs. To prove this idea (i.e., the validity of the determined thresholds for other languages without human validation), we used this system for French-English and Persian-English language pairs as well. We randomly extracted 1000 pairs from the Normal distribution over Fr-En and Fa-En pairs and used Google translate service to generate the translations of these sentences in English. Then, we computed the BLEU scores of the English sentences in parallel sentences and compared them to the Normal curve of BLEU scores in Section 6.1. Since the translation service is the same, the computed BLEU scores are comparable. We observed that 98% of French-English and 96% of Persian-English sentence pairs, which fall in $3\sigma^+$ and $2\sigma^+$ regions of the model probability curve, are in the range of $2\sigma^+$ and $3\sigma^+$ area of the BLEU score Normal curve, too. This gives us an estimate of the quality of parallel data as well. As we discussed earlier, we are interested in perfect parallel sentences which are clustered in the highest σ regions. However partial parallel sentences in lower regions can be used for certain purposes too.

As we see in Table 3 there is a high correlation between the scores generated by the model and the BLEU scores of the parallel sentences. Therefore, we argue that irrespective of the language, the model is capable of extracting parallel sentences from any available language pair in Wikipedia with at least 95% accuracy in the last two σ regions. For other σ regions, although the confidence levels and respective accuracies are lower than the higher regions, partial parallel sentences still can be identified. In Figure 3 some parallel sentences with their probabilities estimated by the model are presented.

At the end to compare our results with other similar works, we applied our model on the German-English dataset of BUCC 2017 second shared task on parallel sentence extraction from comparable corpora (Zweigenbaum et al., 2017) and we obtained a state-of-the-art result on it. The results are presented in Table 4

Model	**Precision**	**Recall**	**F1**
VIC (Azpeitia et al., 2017)	88%	80%	84%
This work	**89%**	**83%**	**86%**

Table 4: System results on the German-English dataset of the second shared task of BUCC 2017

7 Error Analysis

As a qualitative study and a complement to the second task, we asked our evaluators to mention a reason why they think a pair of sentences might not be parallel. Based on a short data inquiry, we provided the evaluators with five major error types. We asked them to expand the list of errors if none of the provided error types explains why a given sentence pair is not parallel. They added other three errors to our list. These reasons are listed in Box 1. We can categorize these errors in noncritical, neutral and critical categories.

Sentences with noncritical errors such as error 4, 6 and 8 have slight problems and can be considered parallel in some cases. However, to enhance the quality of parallel sentences, these sentences are excluded from the system output. Neutral errors like errors 5 and 7 do not lead to a significant decrease in system performance, while critical mistakes like errors 1, 2, and 3, cause severe system malfunctioning.

Figure 3: Parallel Sentences and their model generated probabilities. The two boxes in the first row are gold parallel sentences translated by our translators. We returned the sentences to their original documents and used our model to estimate their probabilities. The two boxes in the second and third rows are test English-French and English-Persian sentences, respectively.

In the following, we analyze each of these categories in detail.

1. Errors 4 and 6 are opposite of each other: either source or target sentence contains more information than its counterpart. It is mainly caused by translators adding information to the translated version. In case 8, there are minor discrepancies between source and target sentences. Some of these cases can be considered parallel pairs though. The majority of errors in $3\sigma^+$ and $2\sigma^+$ belong to this category.

2. In case 5, a semi-parallel sentence is detected as a parallel sentence. However, the two sentences do not have enough semantic overlap and are therefore not parallel. In case 7, despite dealing with the same topic, sentence pairs have diverging contexts and settings and are consequently not parallel. We found a few errors of this type only in the $1\sigma^+$ region.

3. In case 1, the source and target sentences are different with no semantic overlap. In case 2, the source or target sentence is incomplete, which is caused by a malfunctioning sentence segmenter. In case 3, extracted source and target sentences are in the same language, which occasionally happens when

- for whatever reason - some sentences are not translated in the target text. The majority of errors in $3\sigma^-$, $2\sigma^-$, and $1\sigma^-$ belong to this category.

The errors in the first category require more elaborate linguistic analysis to be eradicated than the errors in the third category. The errors in the second category are barely detrimental to the system performance.

Default Reasons:
1- Totally different sentences
2- Incomplete sentences
3- Source and target sentences are in the same language.
4- More than half of the meaning is conveyed but not parallel.
5- Less than half of the meaning is conveyed.

Added Reasons:
6- Target sentence is correct but contains more information than the source sentence.
7- Information incorrect/different
8- Small detail(s) missing

Box 1: Human-judged reasons for lacking parallelity of extracted sentence pairs. Although items 4, 6, and 8 are strong candidates for parallel sentences, they are excluded from parallel sentence extraction to increase overall quality. Items 5 and 7 are trivial to the system performance. Items 1, 2, and 3 are serious system errors.

8 Conclusion

We introduced a language-independent parallel sentence extractor using an end-to-end deep neural network architecture. Our system extracts parallel sentences form comparable pages in Wikipedia. Using the gold parallel data compiled by our human evaluators for the German-English language pair, we showed that the system is highly accurate in extracting parallel sentences in other languages as well. Using the system thresholds estimated by human evaluation, we extracted high-quality parallel sentences for German-English, French-English, and Persian-English language pairs. Our model also obtained a state-of-the-art result on the German-English dataset of BUCC 2017 second shared task. In future work, we aim to improve the system by eradicating its errors and performing the translation step seamlessly without the need for any external translation services[3].

Acknowledgments

This work is part of the project ”TransBank: A Meta-Corpus for Translation Research”, funded by the Austrian Academy of Sciences (grant number GD 2016/56). The author thanks Michael Ustaszewski and Andy Stauder from TransBank group at the University of Innsbruck for their thoughtful comments on the final draft. He also appreciates Esther May Rathbone and Carolyn R. Atzl for their participation in human evaluation task. The computational results presented in this work have been achieved (in part) using the HPC infrastructure LEO of the University of Innsbruck.

[3]The data of this work are available at https://github.com/ExperimentalTransbank/WikiParal

References

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the EACL Workshop on New Text*.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martinez Garcia. 2017. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora. Association for Computational Linguistics (ACL)*.

Luciano Barbosa, Vivek Kumar Rangarajan Sridhar, Mahsa Yarmohammadi, and Srinivas Bangalore. 2012. Harvesting parallel text in multiple languages with limited supervision. In *Proceedings of The Conference on Computational Linguistics (COLING)*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a chinese–japanese parallel corpus from wikipedia. In *Proceedings of the 9th Conference on International Language Resources and Evaluation*.

Mark W. Davis and Ted Dunning. 1995. A trec evaluation of query translation methods for multi-lingual text retrieval. In *Proceedings of Text Retrieval Conference (TREC)*.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*.

Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2014. Attentive pooling networks. *arXiv:1602.03609v1*.

Jeffry L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2).

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: a study and an open task. In *Proceedings of IEEE ASRU Workshop*.

Pascale Fung and Kenneth Ward Church. 1994. Kvec: A new approach for aligning parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*.

I. Dan Melamed. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*.

Mehdi Mohammadi and Nasser GhasemAghaee. 2010. Building bilingual parallel corpora based on wikipedia. In *Proceedings of International Conference on Computer Engineering and Applications*.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Lecture Notes in Computer Science*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4).

Douglas W. Oard. 1997. Cross-language text retrieval research in the usa. In *Proceedings of the Third DELOS Workshop on Cross-Language Information Retrieval*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 1993. Bleu: a method for automatic evaluation of machine translation. *Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rico Sennrich and Martin Volk. 2011. Iterative, mt-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (COLING)*.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor Tron. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*.

Dekai Wu. 1994. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. 2006. Automatic acquisition of chinese-english parallel corpus from the web. In *Proceedings of the 28th European Conference on Information Retrieval*.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop Building Using Comparable Corpora*.

Dan Ştefănescu and Radu Ion. 2013. Parallel-wiki: A collection of parallel sentences extracted from wikipedia. In *Proceedings of the 14th Conference on Intelligent Text Processing and Computational Linguistics*.

Adéla Štromajerová, Vít Baisa, and Marek Blahuš. 2016. Between comparable and parallel: English-czech corpus from wikipedia. In *Proceedings of Advances in Slavonic Natural Language Processing*.