SARAL: A Low-Resource Cross-Lingual Domain-Focused Information Retrieval System for Effective Rapid Document Triage

Elizabeth Boschee◇, Joel Barry◇, Jayadev Billa◇, Marjorie Freedman◇, Thamme Gowda◇, Constantine Lignos◇, Chester Palen-Michel◇, Michael Pust◇, Banriskhem K. Khonglah♣, Srikanth Madikeri♣, Jonathan May◇, Scott Miller◇

◇Information Sciences Institute, University of Southern California

♣Idiap Research Institute

{boschee,joelb,jbilla,mrf,tg,lignos,cpm,pust,
jonmay,smiller}@isi.edu

{banriskhem.khonglah,srikanth.madikeri}@idiap.ch

Abstract

With the increasing democratization of electronic media, vast information resources are available in less-frequently-taught languages such as Swahili or Somali. That information, which may be crucially important and not available elsewhere, can be difficult for monolingual English speakers to effectively access. In this paper we present SARAL, an end-to-end cross-lingual information retrieval (CLIR) and summarization system for low-resource languages that 1) enables English speakers to search foreign language repositories of text and audio using English queries, 2) summarizes the retrieved documents in English with respect to a particular information need, and 3) provides complete transcriptions and translations as needed. The SARAL system achieved the top end-to-end performance in the most recent IARPA MATERIAL CLIR+summarization evaluations.

1 Introduction

The task of searching for a needle of relevant information in a haystack of documents is not as daunting as in previous eras, thanks to decades of information retrieval research progress. Most of us engage in this behavior daily when we search the web. Powerful IR algorithms choose the most likely matches for our queries, but humans also play a crucial role: we are typically presented with a list of ranked results, accompanied by small snippets of relevant content, and we make the final decision with this information in hand.

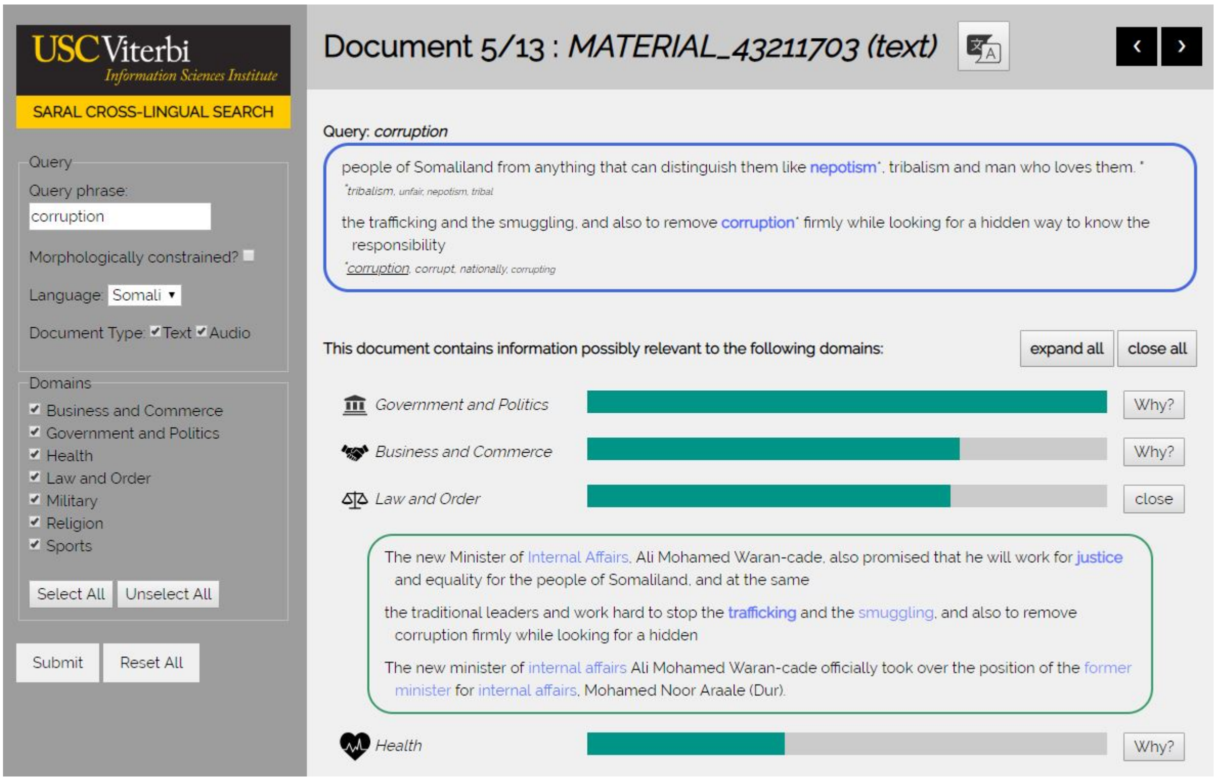
Unfortunately, when the information content is in a language the searcher does not understand, serious challenges can arise. This is the problem of cross-lingual information retrieval (CLIR), and there are several straightforward approaches to this problem, many of which have been well-studied. One can translate queries into the language of the search corpus before matching, or conversely translate the documents into the language of the query. Both approaches naturally rely on the availability of good-quality translation, which improves as more parallel data is available. Thus, CLIR may be adequate when the languages are English, French, Spanish, etc., but will be less effective for lower-resourced languages such as Swahili or Somali.

Moreover, the crucial role played by humans in triaging results is complicated in a low-resource cross-lingual setting, since the system must somehow present the user with the context for its retrieval, e.g. an English speaker with the context for a Swahili document. But if the quality of the machine translation (MT) is too poor, just showing the surrounding text (à la Google) will be insufficiently helpful. This problem is exacerbated when the original source is audio transcribed by a low-resource automatic speech recognition (ASR) model, since ASR errors will propagate through MT.

In this paper we present SARAL (**S**ummarization and domain-**A**daptive **R**etrieval **A**cross **L**anguages[1]), an end-to-end system that addresses these challenges. SARAL operates over both text and audio input documents from a diverse set of genres (e.g. news, conversational speech, etc.), answering user queries by summarizing the retrieved documents *in English* with respect to a user's particular information need. Requests can be expressed as a combination of a query phrase (e.g. *foreign investments*) and a set of one or more desired document domains (e.g. *Health* or *Military*). The SARAL system achieved the top end-to-end performance in the most recent CLIR+summarization evaluations conducted by

[1] SARAL (सरल) is a Hindi word which can be translated as *ingenious* or *simple*, depending on the relevant context.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 19–24
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics



Query
Query phrase: corruption
Morphologically constrained? □
Language: Somali ▾
Document Type: ☑ Text ☑ Audio

Domains
☑ Business and Commerce
☑ Government and Politics
☑ Health
☑ Law and Order
☑ Military
☑ Religion
☑ Sports

Document 5/13 : *MATERIAL_43211703 (text)*

Query: *corruption*

people of Somaliland from anything that can distinguish them like nepotism*, tribalism and man who loves them. "
*tribalism, unfair, nepotism, tribal

the trafficking and the smuggling, and also to remove corruption* firmly while looking for a hidden way to know the responsibility
*corruption, corrupt, nationally, corrupting

This document contains information possibly relevant to the following domains:

Government and Politics

Business and Commerce

Law and Order

The new Minister of Internal Affairs, Ali Mohamed Waran-cade, also promised that he will work for justice and equality for the people of Somaliland, and at the same

the traditional leaders and work hard to stop the trafficking and the smuggling, and also to remove corruption firmly while looking for a hidden

The new minister of internal affairs Ali Mohamed Waran-cade officially took over the position of the former minister for internal affairs, Mohamed Noor Araale (Dur).

Health

Figure 1: The SARAL cross-lingual search interface, which returns English query-focused snippets, domain relevance confidence backed up by domain snippets, and full-text transcription (where relevant) and translation.

the IARPA MATERIAL program.

The contributions of this paper are:

1. SEARCHER, a novel CLIR approach designed for low-resource conditions that relies on the construction of a shared semantic space learned from bitext and monolingual corpora
2. An intuitive snippet extraction and presentation design which has been shown in human studies to provide readers with sufficient evidence to filter out erroneous query matches and preserve good ones, even in low-resource conditions
3. The entire operable SARAL system itself, an end-to-end CLIR and summarization system that combines SEARCHER and traditional IR techniques and applies them to text and speech documents in low-resource languages

An example of the user interface is shown in Figure 1. An instance of the system with Swahili and Somali data may be queried at https://material.isi.edu (register with token *PpnOMgavHR3j*). A short demonstration video is also available.[2]

2 SARAL System Overview

2.1 Automatic Speech Recognition

We transcribe audio data using two systems developed for SARAL by Idiap and ISI. The Idiap system trains 3 Kaldi-based LF-MMI models with a CNN-BLSTM architecture, with targets derived from alignments produced by HMM/GMM models. The first model is trained with standard data augmented by perturbing audio speeds, the second with data augmented by adding noise and then speed perturbation, and the third with bottleneck features extracted from a multilingual system (Tagalog, Swahili, Zulu, Turkish and Somali). The three systems are then fused by stacking lattices and minimum Bayes Risk (MBR) rescoring. The ISI system uses eight Kaldi-based end-to-end LF-MMI trained TDNN-F grapheme acoustic models. Audio data is decoded with each of the models with a trigram LM, followed by rescoring with an RNN-LM to generate lattices. Similar to the Idiap system, the final transcript is generated by stacking lattices from these models, followed by MBR rescoring on the composite lattice.

Based on performance on a development set, we use the Idiap system for conversational speech and

[2] https://youtu.be/TslZiwPejcU

the ISI system for topical and news broadcasts. All models are trained with 40 hours of the transcribed audio provided in the MATERIAL program, as well as ~500hrs of YouTube data used for unsupervised training. For Somali, language models use ~320M words, primarily composed of webcrawl data (~230M words) and the so16 Somali Web Corpus (~70M words); for Swahili, they use ~100M words of webcrawl data. For comparison, a high-resource language would typically be trained with thousands of hours of speech and a language model generated from more than a billion words of data.

2.2 Machine Translation

Our low-resource MT architecture is a system combination (Heafield and Lavie, 2010) of a Transformer-based neural model (Vaswani et al., 2017) and a statistical syntax-based model (Galley et al., 2006), which bring complementary strengths, particularly in low-resource conditions. All models are trained with fewer than 2M words of parallel data.[3] By contrast, in the WMT 2018 shared task (Bojar et al., 2018) most language pairs had 4M or more words, and many had more than 10M words. To further adapt to low-resource conditions, we augment our neural system with 14.5M words of crawled English region-relevant data with parallel Somali or Swahili obtained from backtranslation Transformer models (Sennrich et al., 2016a). Transformer model hyperparameters are "out-of-the-box" except that the shared Byte Pair Encoding (Sennrich et al., 2016b) vocabulary is set to approximately 8,000.

2.3 Cross-Lingual Information Retrieval

We employ a combination of two approaches to cross-lingual information retrieval. The first relies on term-level matching in both the original document and its machine translation(s). Source-language matching is mediated via translation tables derived from the word alignments used by our syntax-based MT system. Terms are expanded using transformations of varying expected accuracy, e.g. stemming, WordNet transformations (Fellbaum, 1998), paraphrases (Pavlick et al., 2015), semantic similarity (Huang et al., 2018), and combinations of the above. For multiword search strings, all terms must match in the same sentence, but not necessarily in the same translation or even the same language. For instance, the Somali phrase *xilli roobaadka* could be translated *rainy season* or *rainy time*. An English-only search for *rainy season* might miss a translation that reported only *rainy time*. However, our hybrid search will match *rainy* in English and *xilli* in Somali, allowing for a match for the phrase across the two languages.

Our second approach, SEARCHER (**S**hared **E**mbedding **ARCH**itecture for **E**ffective **R**etrieval), maps both queries and documents into a shared embedding space and performs retrieval there, rather than relying on translation of either the document or the query terms. However, during development, we found that standard cross-lingual embeddings derived from monolingual corpora, even when aligned using sophisticated transformation techniques (e.g. Lample et al., 2018), did not provide the @1-precision necessary for the specific requirements of MATERIAL's "lexical" queries, where only documents containing precise translations of query terms are judged responsive.

To obtain sufficient precision, we train a proxy task based on sentence relevancy. Here, a sentence S is considered responsive to a query q if at least one plausible translation of S contains the term q. Training samples are derived from parallel corpora. Sample queries are drawn from the English side, with their corresponding foreign-language sentences as positive examples and other randomly-drawn foreign-language sentences as negative examples. The SEARCHER model consists of a convolutional encoder (similar to Gehring et al. 2017) for encoding foreign-language sentences, a query embedding matrix, an attention mechanism for aligning query terms with specific foreign-language terms, and a matching network to determine relevance. The model was optimized using a cross-entropy objective. In recent experiments, SEARCHER's performance exceeded that of the term-level matching approach, improving AQWV (see Section 3) from 23.1 to 25.2 on the Somali MATERIAL evaluation corpus, even when translation is performed by state-of-the-art MT systems.

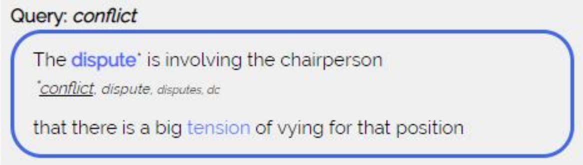
2.4 Domain Identification

The New York Times Annotated corpus[4] provides ~2M articles with topic annotations from a closed topic set. For each domain of interest,[5] we manu-

[3] Data was provided by the IARPA MATERIAL program and by LDC as part of the DARPA LORELEI program (Somali: LDC2016E91; Swahili: LDC2017E64).

[4] `https://catalog.ldc.upenn.edu/LDC2008T19`

[5] Business & Commerce, Government & Politics, Health, Law & Order, Military, Religion, Sports



Query: *conflict*

The dispute* is involving the chairperson

**conflict*, *dispute*, *disputes*, *dc*

that there is a big tension of vying for that position

Figure 2: Example summary for the query *conflict*.

ally select the topics that best map to the domain, giving us a set of in-domain documents. We then calculate a score for each n-gram ($n \leq 3$) that represents how indicative it is of a particular domain, simply $count_{in\_domain}/count_{all}$. We discard all n-grams involving capitalized letters (mostly names) as likely irrelevant (or even misleading) to the target datasets (e.g. Somali news). Our binary domain classifier then has three parameters: a threshold for unigrams, a threshold for bi/trigrams, and the number of n-grams whose scores meet those thresholds that must be found for a document to be considered in-domain. We tune these parameters for each domain via grid search on the development corpus, optimizing for AQWV on the CLIR task.

2.5 Summary Generation

The goal of summarization is to concisely explain, in English, a particular document's relevance to a query. Our primary approach highlights in blue those terms ranked most highly by our CLIR and displays them in a fixed-context window. Semantically related words are colored in lighter blue, as with *tension* in Figure 2. When query terms are found in the source language or matched in the SEARCHER embedding space, we attempt to highlight aligned terms in one of our English machine translations, where possible. (In some cases, no translation of a particular foreign term might be found; in that case we simply present the whole sentence without highlighting.)

The primary barrier to providing accurate summaries is poor MT quality. Even if an exact match is highlighted, the context may be so garbled that a reader is unable to label it as a reliably relevant match. To mitigate this, we provide additional context for the MT system's decisions, specifically the set of options the system considers when producing word(s) matching the query. For instance, consider a summary for *back injuries*. If the word *back* was translated from the Swahili word *mgongo*, we might show alternate translations *spine*, *backbone*, and *spinal*, reassuring the reader that the translation of *back* is correct and of the appropriate word sense. In contrast, if the word was originally translated from *kurejea*, we would present alternative translations *return*, *returning*, *referring*, leading the reader to correctly identify a false alarm.

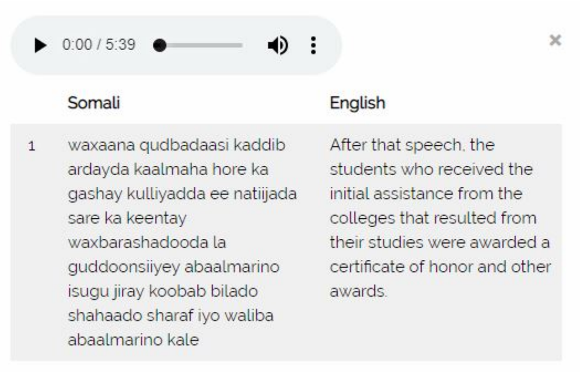
For the purposes of summarization, we provide this kind of information via footnotes (see Figure 2), where the size of a word in the footnote reflects how likely the system thinks it is (in isolation) to be a translation of the original source term. We also underline the exact query term if it is present in that list, to help draw the user's attention to it.

We generate summaries for domains using the n-grams extracted for domain classification (Section 2.4). We identify these n-grams in an English machine translation of a document and create multiple candidate display windows of varying size for each. We then employ a greedy search to select and merge such windows to (a) include as much domain-relevant information as possible (a function of both the number of domain-relevant terms and their quality), (b) present exactly as much context as is necessary to make the terms understandable, and (c) avoid redundancy / prefer diversity. When presenting summaries to the user, we highlight domain-relevant terms in blue, with the shade intensity indicating the strength of its relevance to the domain. A sample summary for the Law and Order domain is shown in Figure 1.

2.6 User Interface Design

SARAL's user interface allows users to search for a single English query phrase. Following the most common practice of the MATERIAL program, we focus on direct cross-lingual search rather than conceptual expansion. So, for the query *vaccine*, synonyms (e.g. *immunization*) and morphological variations (e.g. *vaccinated*) would be considered responsive, but a sentence generically discussing *methods for the prevention of the flu* would not. (Users may also opt to exclude morphological variations.) Users also select the target language and optionally restrict to either text or audio documents.

In the MATERIAL program, queries typically require exactly one domain. However, a user's interests might extend to more than one domain at a time. We therefore allow the user to select multiple domains; any document that matches at least one domain of interest is allowed to be returned as relevant. To avoid crowding the screen when a document is relevant to multiple domains, we show instead, for each document, a bar graph displaying



0:00 / 5:39

	Somali	English
1	waxaana qudbadaasi kaddib ardayda kaalmaha hore ka gashay kulliyadda ee natiijada sare ka keentay waxbarashadooda la guddoonsiiyey abaalmarino isugu jiray koobab bilado shahaado sharaf iyo waliba abaalmarino kale	After that speech, the students who received the initial assistance from the colleges that resulted from their studies were awarded a certificate of honor and other awards.

Figure 3: ASR & MT excerpt for an audio document.

the relative strength of each domain that the system identified as being potentially represented in a document. Clicking on the *Why?* button next to a domain displays the evidence that the system found for that domain, i.e. the domain-specific summary, as shown in Figure 1.

For the purposes of the demonstration, we restrict query summaries to 50 words, keeping them comfortably at the top of the page and quickly gistable. We allow 80 words for each domain summary, enough to provide convincing evidence without being too verbose to skim quickly.

Finally, we provide full access to each source document (original text or audio; if audio, we also provide the automatically-generated transcription) and an English machine translation, for the user who wants to dig deeper into the context of a response. A small excerpt is shown in Figure 3.

2.7 New Languages

It is simple to add a new language to the system. In a recent exercise, we brought up an end-to-end system in Lithuanian in three days using the speech and parallel text resources provided by the MATERIAL program; this required only a few hours of actual human effort. The two largest bottlenecks for improved performance over the three-day system are data collection (scraping monolingual data from the web to improve ASR language models) and ASR model training. With ten days, we were able to bring up a significantly improved ASR system in Lithuanian; with more efficient use of compute resources (e.g. parallelizing the web scraping), this time could be significantly reduced.

3 System Evaluation & Analysis

The Phase 1 MATERIAL evaluation was performed on a corpus of ∼15K Somali documents annotated for relevance for 1,000 queries by native speakers. The official evaluation metric is AQWV (Average Query Weighted Value),[6] which uses a parameter β to balance missed detections and false alarms.

End-to-end AQWV was calculated after human readers triaged an initial set of system results, removing those documents they judged to be false alarms using *only* the English summaries generated by the system. Documents were sampled evenly across queries and across true positives and false alarms; system performance was then projected to any unassessed documents. For the SARAL system, ∼15K query/document summaries were assessed, using Amazon Mechanical Turk. Overall, the SARAL system was the top-ranked end-to-end system in the evaluation.[7]

Projected across all responses, the SARAL summarization component results in the acceptance of 87% of true positives and the rejection of 45% of false negatives. Rates are essentially consistent across speech and text documents. Because the AQWV β for the evaluation penalizes misses much more than false alarms, these results are consistent with our goal of minimizing false rejections even if that means retaining more false positives.

The majority of errors on true positive documents come from insufficient summaries. For instance, a query about *deception* results in the summary text *Punamin was arrested for trafficking, but he made amazing* ***cheating*** *that he thought about the long arrest*. Two alternative translations provided for cheating are *deception* and *trick*. Still, the English context is difficult to understand. Thus although it is in reality a true positive, it is not unreasonable that a human rejected it.

Human acceptance of a false positive happens most frequently when readers accept an alternate translation as accurate when the context did not make sense. For instance, a query for *midwife* returns summary text *I would like to advise you to be united people who create their own* ***skills*** *... you will be a company that will support themselves*. Our system indicates that an alternate translation for *skills* could be *midwife*, which is accepted by the reader even though clearly incorrect in context.

A so-called false positive found by the system—

[6] https://www.nist.gov/sites/default/files/documents/2019/04/02/openclir19_evalplan_v1.19.pdf

[7] Four end-to-end systems were evaluated from four separate teams. However, per program policy, only system rank may be publicly reported, so we cannot provide any further details on cross-system comparisons here.

and retained by human readers during triage—can actually be a true positive that was missed by the original foreign-language annotator. For instance, a query for *mockery* returns *will present a exhibition to show* ***insults*** *to our Prophet ... aimed at presenting images of* ***insulting*** *Prophet Muhammed*. It seems reasonable that *insults* here is a translation variant for *mockery*; both our system and a human reader think so. This shows the strength of the system; not only can it provide a monolingual speaker with access to content in low-resource foreign languages, but it can sometimes surpass search by native speakers.

4 Related Work

Recent research in CLIR and query-based summarization uses expansive, concept-based definitions of relevance. For example, given the query *agriculture*, documents are relevant if they describe fields, pastures, or crops, even if the word *agriculture* is not used, and the goal of summarization is to show that the document as a whole is relevant. In contrast, in this work we aim to retrieve documents that meet a more precise notion of relevance, similar to that used for keyword spotting. This goal influences our retrieval approach, which seeks to account for variation in translation but does not perform more expansive embedding-based query expansion, and the summarization approach, which presents in-context search term matches rather than a narrative summary of the document as a whole.

5 Conclusion

The SARAL system provides a monolingual user with effective access to multimodal information in lower-resourced languages through a user interface that enables rapid triage of system results. We look forward to future work improving the quality of the underlying components for low-resource settings as well as expanding the user interface to incorporate additional semantic constraints or requests.

Acknowledgments

Thanks to Heng Ji for fruitful discussions. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract # FA8650-17-C-9116. 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 ODNI, 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.

References

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proc. WMT*, Belgium, Brussels.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. COLING/ACL*, Sydney, Australia.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proc. ICML*, Sydney, Australia.

Kenneth Heafield and Alon Lavie. 2010. Voting on n-grams for machine translation system combination. In *Proc. AMTA*, Denver, Colorado, USA.

Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Multi-lingual common semantic space construction via cluster-consistent word embedding. In *Proc. EMNLP*, Brussels, Belgium.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proc. ICLR*, Vancouver, Canada.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proc. ACL*, Beijing, China.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proc. ACL*, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proc. ACL*, Berlin, Germany.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*.