Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop, pages 22–25
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

SEARCHER: Shared Embedding Architecture for Effective Retrieval

Joel Barry, Elizabeth Boschee, Marjorie Freedman, Scott Miller
Information Sciences Institute, University of Southern California
{joelb, boschee, mrf, smiller}@isi.edu

Abstract

We describe an approach to cross lingual information retrieval that does not rely on explicit translation of either document or query terms. Instead, both queries and documents are mapped into a shared embedding space where retrieval is performed. We discuss potential advantages of the approach in handling polysemy and synonymy. We present a method for training the model, and give details of the model implementation. We present experimental results for two cases: Somali-English and Bulgarian-English CLIR.

Keywords: CLIR, cross-lingual embeddings

1. Introduction

A fundamental design decision in cross-lingual information retrieval is whether to translate the queries, the documents, or both. In this paper, we discuss a substantially different alternative where neither the query nor the document is translated. Instead, both the queries and documents are projected into a shared embedding space and retrieval is performed there. The approach offers potential advantages in handling synonymy, i.e. where synonymous query terms can match a single document term (or vice-versa), as well as for document-language polysemy, i.e. where a particular document term can have one of several meanings depending on context. In tests on two languages, Somali and Bulgarian, we observed a level of performance that is competitive with the "document translation" approach, including when translation is performed using a state-of-the-art tensor-to-tensor model. For one of the languages, Somali, the shared embedding approach was also able to outperform a hybrid strategy involving both query and document translation. All experimental results were from IARPA's MATERIAL evaluation task.

2. Initial Experiments

Methods for constructing cross-lingual (and multilingual) word embeddings have been extensively investigated for the past several years (Hermann and Blunsom, 2014; Luong, Pham, and Manning, 2015; Gouws, Bengio, and Corrado, 2015) and several pre-trained resources are publicly available. To begin exploring the possibility of applying shared embeddings for CLIR, we constructed a baseline system and tested a few state-of-the-art publicly-available variants, including MUSE (Conneau et al., 2017). The baseline system architecture is shown in Figure 1.

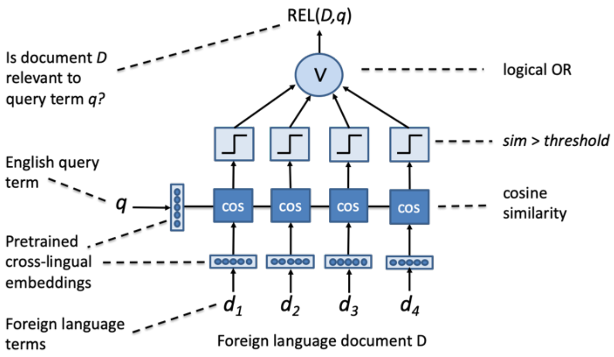


Figure 1: Baseline architecture

In this system, document relevancy is determined based on cosine distance between query and document terms. More specifically, a document is considered responsive to a query if at least one of the document words is within a fixed threshold (in embedding space) of the query. Despite basing our experiments on state-of-the-art embeddings, initial performance was low. The AQWV score (Actual Query Weighted Value) for MATERIAL's Swahili-English analysis set was 0.03; for Tagalog-English it was 0.07.

2.1 Limitations of the Baseline Approach

Three factors seemed to account for the low AQWV scores. First, embedding spaces are not uniform; some regions are densely packed with words while other regions are only sparsely populated. Thus, no consistent interpretation of distance exists, making the selection of a single matching threshold problematic. Second, although simple linear transformations are capable of aligning semantically-related words across languages, the alignments are not sufficiently precise to identify exact term translations – particularly for MATERIAL's lexical queries. Finally, our retrieval mechanism was massively under-parameterized; initial experiments attempted to optimize a complex CLIR task by adjusting only a single scalar threshold parameter.

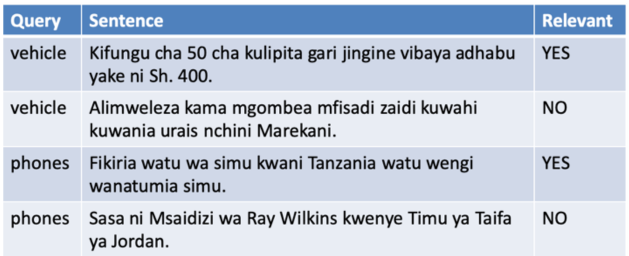
3. Training Data and Objective Function

Overcoming these limitations would require a sufficiently parameterized model that could be trained for the CLIR task. Implicit in this approach is the need for training data and for a well-defined training objective. In principle, data provided by the MATERIAL program could provide the training examples and AQWV could serve as the objective function. However, MATERIAL's rules explicitly prohibit directly training on this data and, in any case, the relatively small number of queries and relevance judgements is insufficient to train an adequate model (e.g., embedding parameters alone require estimating millions of floating-point values).

Instead, we defined a simplified sentence-retrieval task for which training data is readily available. Specifically, given an English query term (q) and a foreign language sentence (S):

- **Sentence S is relevant to query q if there exists at least one plausible translation of S containing q.**

For this proxy task, large numbers of training examples can be extracted from a parallel corpus such as used to train machine translation systems. Specifically, any English term that occurs anywhere in a bitext sentence can be treated as a query and its corresponding foreign-language sentence treated as a positive example. Negative examples can be randomly drawn from foreign-language bitext sentences (any randomly selected sentence is probably not relevant, but we can additionally verify that its corresponding English sentence does not contain the query term).



Query	Sentence	Relevant
vehicle	Kifungu cha 50 cha kulipita gari jingine vibaya adhabu yake ni Sh. 400.	YES
vehicle	Alimweleza kama mgombea mfisadi zaidi kuwahi kuwania urais nchini Marekani.	NO
phones	Fikiria watu wa simu kwani Tanzania watu wengi wanatumia simu.	YES
phones	Sasa ni Msaidizi wa Ray Wilkins kwenye Timu ya Taifa ya Jordan.	NO

Figure 2: Examples of training instances

Figure 2 shows examples of training instances from a Swahili/English parallel corpus. The sentence in the first row translates to "The fine for passing another vehicle improperly is 400 shillings." Similarly, the sentence in the third row translates to "Think about people with phones since in Tanzania so many people are using phones." The sentences in rows 2 and 4 are randomly selected Swahili sentences that do not contain the query term.

Given a training corpus of such examples, the probability that a sentence S is relevant to a query q, i.e. $P(rel|S,q)$, can be optimized using the standard cross-entropy objective function H

$$H(X) = \sum_{X} z * -\log\left(p(rel|S,q) + (1-z)\right.$$
$$\left. * -\log\left(1 - p(rel|S,q)\right)\right)$$

where X is the set of training examples and z are the true labels (1 for relevant, 0 for irrelevant).

For the actual MATERIAL task, the relevance of a document to a query phrase is taken as the maximum relevance over sentences in the document.

4. Model Architecture

Now that we have identified suitable training data and an objective function, we next consider the challenge of model design.

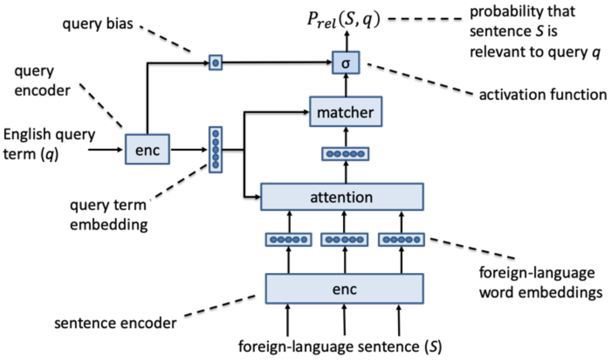


Figure 3, Generic SEARCHER Architecture

Here we introduce the following elements:

- Query encoder: maps English terms into the shared embedding space
- Sentence encoder: maps foreign-language terms into the shared embedding space
- Attention mechanism: selects regions of the sentence based on the query
- Matching mechanism: determines how closely the selected region matches the query
- Activation function: maps matching scores to probability values

An overview of the generic SEARCHER architecture is shown in Figure 3. The retrieval process proceeds as follows. First, each foreign-language word is mapped into the shared embedding space. These embeddings are contextualized, as described in Sections 5 and 6. Next, the English query term is mapped into the common embedding space. An attention mechanism then selects the region of the foreign-language sentence that appears most relevant to the query and outputs its embedding. The selected region's embedding is compared to the query by a matching function which outputs a matching score. Finally, the matching score is passed through an activation function that produces the probability of relevance. Importantly, this activation function also receives a separate query-specific bias value. This bias value helps overcome non-uniformity in the embedding space by requiring some terms to match more closely than others depending on the density of their surrounding neighborhoods. In all of our experiments, we use a sigmoidal activation function.

5. Contextualized Embedding Spaces

Beginning with models such as BERT (Devlin et al., 2018) and ELMO (Peters et al., 2018), contextualized embeddings have proven useful for a wide range of tasks. While MATERIAL's queries typically contain only one or a few words, and therefore offer little opportunity for query contextualization, our proxy CLIR task evaluates relevance over complete sentences, offering the possibility of contextualizing document embeddings. A potential advantage of such contextualization is the resolution of polysemous terms. Specifically, a contextualized model can learn to situate polysemous terms in different regions of the embedding space depending on context. For example the Swahili term "nyanya" can be translated alternatively as "grandmother" or "tomatoes," as shown in Figure 4. Ideally, a contextualized model will place the different senses of a polysemous term in different locations in the embedding space, thereby reducing the possibility of spurious matches (e.g. retrieving grandmothers when searching for tomatoes).

We note that in SEARCHER, contextualized embeddings are used only for document terms; non-contextual embeddings are used for query terms.

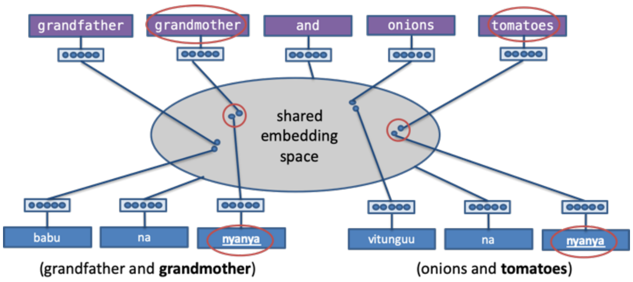


Figure 4: Polysemy in shared embedding space

Performing retrieval in a shared embedding space is also potentially useful for resolving synonymous terms. For example, the Swahili term ‘gari’ can be translated equivalently as “car” or “vehicle,” as shown in Figure 5. Ideally, the model will place synonymous terms in similar positions in the embedding space, thereby increasing the possibility of matching any of the alternatives (e.g. retrieving a document containing “gari” whether the query term is “car” or “vehicle”).

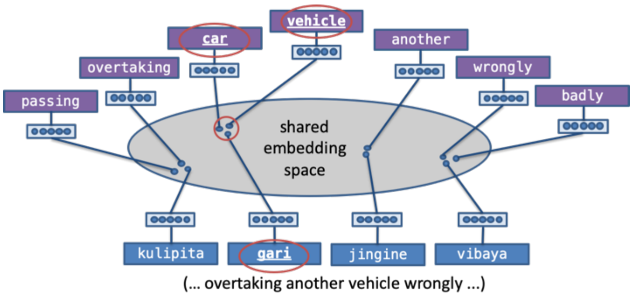


Figure 5: Synonymy in shared embedding space

6. Convolutional Encoder

In this section, we consider details of the sentence encoder mentioned in Section 4. Specifically, SEARCHER’s sentence encoder produces contextualized embeddings using a deep convolutional model consisting of 15 convolution layers, each of diameter 3. This architecture yields a receptive field of 31 words, providing 15 words of context on each side of a term. The encoder is similar to that described in (Gehring et al., 2018).

In detail, each convolution block consists of a dropout layer, a convolution layer, a GLU layer (gated linear units), and residual connections. A fixed embedding size of 512 is maintained throughout the network.

We use an identical encoder in our convolutional machine translation system. In fact, we have found that pretraining the encoder in an MT setting, then transferring the encoder to SEARCHER, and continuing to train the remaining CLIR elements is an effective method for speeding convergence.

7. Simplifications

Our generic SEARCHER architecture leaves room for various alternatives at the level of individual components. For instance, while we use a convolutional sentence encoder, it would be perfectly reasonable to substitute a transformer architecture.

One alternative involving the attention and matching mechanisms leads to a particularly attractive simplification. Specifically, if the attention mechanism is the commonly used form:

$$ATT(S, q) = \sum_{i \in |S|} \alpha_i s_i$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i \in |S|} \exp(e_i)}$$

$$e_i = q \cdot s_i$$

and the matcher is a simple dot product, then the resulting architecture (after some algebra) reduces to that shown in Figure 6.

We have found this simplified architecture to be effective, producing results at least as good as more complex variations. A further simplification is obtainable by replacing the softmax pooling layer with a hard max-pooling layer. Both simplified variations produce similar results. The softmax variation requires fewer training cycles (because max-pooling updates just the single best-matching term on each training cycle, whereas softmax pooling updates all words in proportion to their distance from the query). On the other hand, max pooling appears to yield slightly sharper probability distributions.

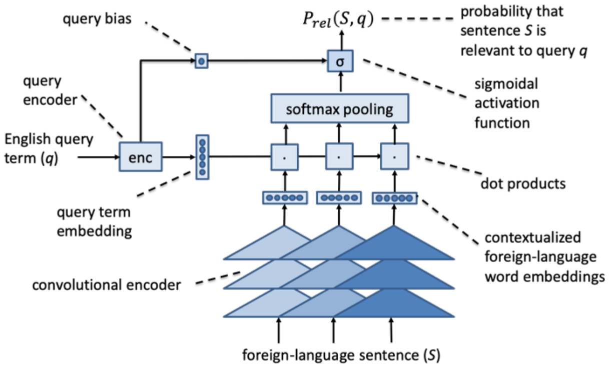

query bias
$P_{rel}(S, q)$
probability that sentence S is relevant to query q
σ
query encoder
sigmoidal activation function
softmax pooling
English query term (q)
enc
·
dot products
query term embedding
contextualized foreign-language word embeddings
convolutional encoder
foreign-language sentence (S)

Figure 6: Simplified SEARCHER Architecture

8. Relation to the Baseline Model

The SEARCHER model shown in Figure 6 bears a striking resemblance to the baseline model described in Section 2. The most important difference is that the SEARCHER model is specifically trained to perform CLIR whereas the baseline model relies on pretrained embeddings. Other differences are:

- Contextualized embeddings replace individual word embeddings
- Dot products replace cosine distances (which are simply normalized dot products)
- Softmax pooling (essentially, a soft OR function) replaces the logical OR
- A sigmoidal activation function (essentially, a soft threshold) replaces hard thresholding
- The positions of the combining function (softmax/logical OR) and the activation function (sigmoid/hard threshold) are exchanged
- A bias term is introduced for each query term

9. Experimental Results

We tested SEARCHER in two MATERIAL languages, Somali and Bulgarian. For each language, we also evaluated traditional translation-based CLIR.

For the Somali case, we compare performance with several different machine translation models. These include syntax-based statistical machine translation and two types of neural machine translation: tensor-to-tensor (Vaswani et al., 2017) and convolutional (Gehring et al., 2018). For the neural models, we follow best practices in training, including the use of substantial back-translated data.

In all cases, the MT system is applied to translate the foreign language documents into English. We also evaluate alternatives where, in addition to translating the documents, we translate the English queries into the foreign language using translation tables obtained by a statistical alignment process. This strategy improves the probability of matching queries to documents by translating in both directions.

Results for Somali, as shown in Table 1, are encouraging. Entries in the table that are designated (+source) indicate the combined strategy where queries are also translated. Evaluating on two different MATERIAL data sets (designated analysis and dev), SEARCHER outperformed the "document translation" strategy for all translation models as well as the combined strategy where both the documents and the queries are translated.

System	AQWV analysis1/q1	AQWV dev1/q1
syntax-based MT	0.1537	0.2110
syntax-based MT + source	0.1643	0.2257
tensor-to-tensor MT	0.1753	0.1852
tensor-to-tensor + source	0.1904	0.2251
convolutional MT	0.1611	0.1965
convolutional MT + source	0.1814	0.2361
SEARCHER	**0.2290**	**0.2502**

Table 1: AQWV of various systems on Somali

For the Bulgarian case, we compare SEARCHER with only our best machine translation model, a tensor-to-tensor model, and evaluate only on MATERIAL analysis documents. Once again, the MT system is applied to translate the foreign-language documents into English. As before, we also evaluate the combined strategy, translating both documents and queries.

Results for Bulgarian are shown in Table 2. In this case, results are somewhat different. In general, performance is much better. SEARCHER's performance matches the "document translation" strategy alone. However, when query translation is added, the combined translation strategy noticeably outperforms SEARCHER. We suspect that part of the explanation for the differences in relative performance is the amount of training data available. Specifically, large quantities of paracrawl data for Bulgarian provide a significant boost in MT accuracy.

System	AQWV analysis1/q1
tensor-to-tensor MT	0.6527
tensor-to-tensor + source	**0.6998**
SEARCHER	0.6546

Table 2: AQWV for Bulgarian

10. Summary

We have conducted numerous experiments with SEARCHER models. We have identified an effective general architecture and derived simplified variations that perform well. We found that training for a proxy task (sentence retrieval) is a useful strategy and that adequate training examples can be derived from bitexts. While much work remains to be done, we have demonstrated that shared embedding space models can be an effective method for CLIR, providing a competitive alternative to document translation models, including those based on state-of-the-art neural MT. In one language, Somali, we found that SEARCHER outperformed all the translation-based alternatives that we evaluated.

11. Acknowledgements

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-C9116. 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.

12. References

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou, (2017). Word Translation Without Parallel Data, arXiv:1710.04087

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, Yann N. Dauphin, (2017). Convolutional Sequence to Sequence Learning, arXiv:1705.03122

Stephan Gouws, Yoshua Bengio, and Greg Corrado, (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In Proc. of ICML

Thang Luong, Hieu Pham, and Christopher D. Manning, (2015). Bilingual word representations with monolingual quality in mind. In Proc. of the Workshop on Vector Space Modeling for NLP.

Karl Moritz Hermann and Phil Blunsom, (2014). Multilingual Models for Compositional Distributional Semantics. In Proc. of ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, (2018). Deep contextualized word representations, arXiv:1802.05365

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, (2017). Attention Is All You Need, arXiv:1706.03762