The 31st Meeting of Computational Lingustics in The Netherlands (CLIN 31)

A Novel Pipeline for Domain Detection and Selecting In-domain Sentences in Machine Translation Systems

Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, Pieter Spronck
Department of Cognitive Science and Artificial Intelligence, Tilburg University, the Netherlands

Introduction

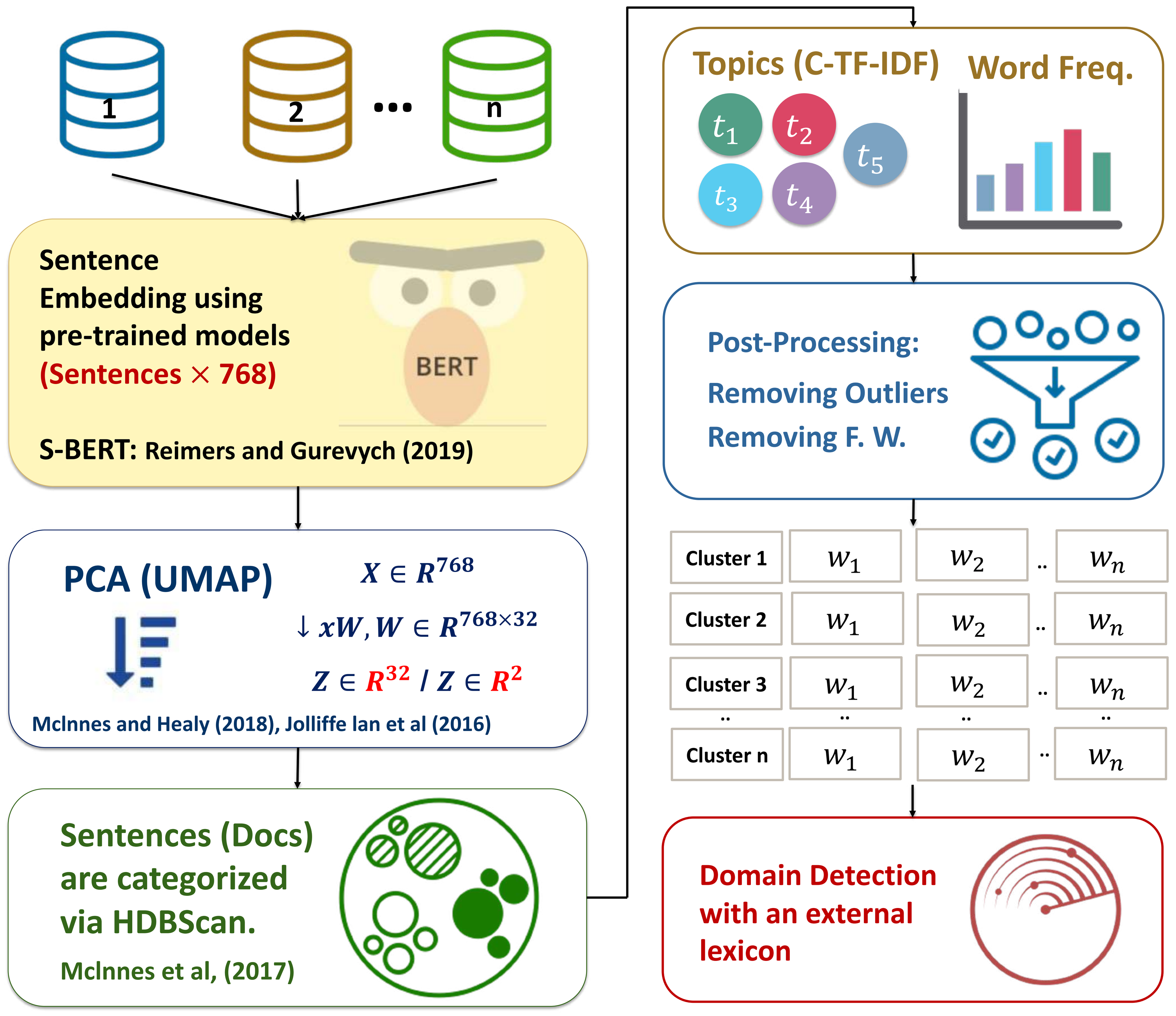
• General-domain corpora are becoming increasingly available for Machine Translation (MT) systems. However, using those that cover the same or comparable domains allow achieving high translation quality of domain-specific MT. It is often the case that domain-specific corpora are scarce and cannot be used in isolation to effectively train (domain-specific) MT systems.

• This work aims to improve domain-specific MT by:

i) A novel semi-supervised pipeline for identifying the distribution of different domains within a corpus, based on domain-specific keyword lists or other lexical resources. These resources are fed into the pipeline and used to identify similar input (i.e., domain-specific) data, within the general domain.

ii) A data selection technique that leverages in-domain monolingual or parallel data to select domain-specific sentences from general corpora according to the distribution defined in (i).

Domain Detection Pipeline



Results

• To test the effectiveness of our approach, the proposed pipeline has been tested on two different domains. i) A **general-domain** corpus called TEP: Tehran English-Persian Parallel Corpus (Pilervar et al 2011) and ii) a monolingual **IT/digital** training data named DeepSentiPers (Sharami et al, 2020)

• We divided the TEP into three samples by shuffling over ~ 600K data such that each one includes about 90K sentences. After domain analysis on subsets, we uncovered each consists of “General” domains more than any other topics. The results shown in Table 1 are based on the top 5-topics for the first sample, without applying the post-processing step.

Topic	Word Freq.	Some words	Domain
1	37712	تو، در، با	Outlier
2	10690	نیست، هیچ، نمی	Outlier
3	4378	زدن، کننده، در	General
4	3815	ممنونم، ممنون	General
5	3593	اشتباه، احمق، بد	General

Table 1: Distribution of domains in TEP

• Table 2 shows results for the DeepSentiPers, a small corpus that consists of 5561 Persian sentences. After analyzing its top n-words, two main domains except for an outlier have been detected.

Topic	Word Freq.	Some words	Domain
1	2655	گوشی، تبلت، صفحه	IT
2	219	خوب، محشره، خاص	General +/- adj.

Table 2: Distribution of domains in DeepSentipers

Discussion

• The output of the last step in the suggested pipeline (domain detection) works based on top n-words. That is, these reveal the most frequent words that occurred within the corresponding cluster. Hence, it would be feasible to extract the top n-words of any other domain-specific corpus. To employ this for data selection, we can select similar sentences based on the matching criteria.

In this use case, a general-domain and domain-specific corpus are fed into the pipeline. Since their top n-words match wrt the defined matching function, the system can distinguish similar sentences from irrelevant ones.

References

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP/IJCNLP.*

McInnes, L., & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv, abs/1802.03426.*

Jolliffe Ian T. and Cadima Jorge 2016 Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A.3742015020220150202

McInnes et al, (2017), hdbscan: Hierarchical density based clustering, Journal of Open Source Software, 2(11), 205, doi:10.21105/joss.00205

Sharami, J.P., Sarabestani, P.A., & Mirroshandel, S.A. (2020). DeepSentiPers: Novel Deep Learning Models Trained Over Proposed Augmented Persian Sentiment Corpus. *ArXiv, abs/2004.05328.*

Pilevar, Mohammad Taher & Faili, Heshaam & Pilevar, Abdol. (2011). TEP: Tehran English-Persian parallel corpus. 6609. 68-79. 10.1007/978-3-642-19437-5_6.

Contact information
Javad Pourmostafa
Email: j.pourmostafa@uvt.nl