EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D4.6: Final cross-lingual news summarisation and visualisation technology (T4.2)

Executive summary
The task of cross-lingual news summarisation and visualisation (Task T4.2) addresses the problem of creating informative condensed textual and visual representations of a larger text or corpus. The proposed solutions address the difficulties of creating these summaries using cross-lingual content in the low-resourced languages of the EMBEDDIA project. We present two technologies for creating cross-lingual textual summaries together with their evaluations where both are shown to compare favourably with the baseline results on the tested low-resourced language summarisation tasks. Finally, we present the visualisations summarising the content of a corpus along temporal and topic axis.

Partner in charge: UEDIN

Project co-funded by the European Commission within Horizon 2020 Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D4.6**
Deliverable full title:	**Final cross-lingual news summarisation and visualisation technology**
Deliverable short title:	**Cross-lingual news summarisation and visualisation**
Document identifier:	**EMBEDDIA-D46-CrosslingualNewsSummarisationAndVisualisation-T42-submitted**
Lead partner short name:	**UEDIN**
Report version:	**submitted**
Report submission date:	**31/10/2021**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Shane Sheehan (UEDIN)**
Co-author(s):	**Saturnino Luz (UEDIN), Jose G. Moreno (ULR), Naveen Saini (ULR)**
Status:	**__ draft, __ final, _x_ submitted**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
05/07/2021	v0.1	Shane Sheehan (UEDIN)	Template.
30/08/2021	v0.2	Jose G. Moreno (ULR), Naveen Saini (ULR)	Draft of text summarisation sections.
30/09/2021	v0.3	Shane Sheehan (UEDIN)	Added TeMoTopic.
05/10/2021	v0.4	Shane Sheehan (UEDIN	Added Abstractive Section.
07/10/2021	v0.5	Shane Sheehan (UEDIN)	Edits to summarisation section.
07/10/2021	v0.6	Shane Sheehan (UEDIN)	Introduction and reduce size of abstractive section.
14/10/2021	v0.7	Shane Sheehan (UEDIN)	Internal review version.
16/10/2021	v0.8	Ravi Shekhar (QMUL)	Internal review.
17/10/2021	v0.9	Marko Robnik-Šikonja (UL)	Internal review.
18/10/2021	v1.0	Jose G. Moreno (ULR)	Applied corrections from internal review.
23/10/2021	v1.1	Shane Sheehan (UEDIN)	Internal review corrections.
25/10/2021	v1.2	Shane Sheehan (UEDIN)	Preparation for quality control.
26/10/2021	v1.3	Jose G. Moreno (ULR)	Added link to repo.
26/10/2021	v1.4	Nada Lavrač (JSI)	Quality checked.
27/10/2021	v1.5	Jose G. Moreno (ULR)	Applied corrections from quality control.
27/10/2021	v2.0	Shane Sheehan (UEDIN)	Quality control changes.
27/10/2021	final	Saturnino Luz (UEDIN)	Final check.
29/10/2021	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

List of abbreviations

CLTS	Cross Language Text Summarisation
ILP	Integer Linear Programming
MEA	Multi-objective Evolutionary Algorithm
CLTS-MEA	Cross Language Text Summarisation - Multi-objective Evolutionary Algorithm
DE	Differential Evolution
CEC	Congress on Evolutionary Computation
IEEE	Institute of Electrical and Electronics Engineers
LSTM	Long Short-Term Memory
MOO	Multi-Objective Optimisation
MBDE	Multi-objective Binary Differential Evolution
ROUGE	REcall-Oriented Understudy for Gisting Evaluation
MSC	Multi-Sentence Compression
CG	Chunk Graph
LDA	Latent Dirichlet Allocation
POS	Part-Of-Speech
SU	Skip Units
CCLTS-MSC	Compresive Cross Language Text Summarisation - Multi-Sentence Compression
ANOVA	ANalysis Of VAriance
MT	Machine Translation
JSON	JavaScript Object Notation

1 Introduction

This deliverable reports on the results achieved in cross-lingual news summarisation and visualisation performed in Task T4.2 of the EMBEDDIA project. This task, which began in M7, is concerned with the development of textual and visual language-independent multi-document news summarisation technologies. The technologies developed in this task enable the summarisation of multilingual news corpora using both contrastive and abstractive techniques along with interactive visual summerization of the corpora content.

In the first deliverable for this task, D4.3, we reported on our initial contrastive cross-lingual text summarisation technique, a visualisation approach to corpus summarisation based on concordance analysis, a graph based visualisation of topics and related terminology which can serve as a visual summary of the topic structure in a corpus, and a prototype temporal topic summarisation visualisation for tracing topic distribution and keywords over time,

In this M33 report, our research on cross-lingual text summarisation and visual summarisation is presented. The work described here does not include work prior to M18, which is described in deliverable D4.3.

In the literature, there are different types of summarisation strategies, based on the type of generated summary: extractive (Litvak & Last, 2013; Saini, Saha, Chakraborty, & Bhattacharyya, 2019; Y. Zhang, Er, Zhao, & Pratama, 2016), abstractive (W. Li & Zhuge, 2019; J. Zhang, Zhou, & Zong, 2016; Duan, Yin, Zhang, Chen, & Luo, 2019; Ladhak, Durmus, Cardie, & McKeown, 2020; Zhu, Zhou, Zhang, & Zong, 2020; Rudra, Goyal, Ganguly, Imran, & Mitra, 2019; K. Yao et al., 2018), and compressive (Linhares Pontes, Huet, Torres-Moreno, & Linhares, 2020). The extractive summarisation technique works by extracting the most relevant sentences from the document. The abstractive summarisation techniques rebuild the contents to generate new sentences. In other words, abstractive summarisation requires the concept of the natural language understanding and generation (Mills & Bourbakis, 2013) to form a summary. Compressive approaches generate a summary by removing non-relevant information from the sentences (Vanetik, Litvak, Churkin, & Last, 2020).

The text summarisation work presented in Section 2 focuses on the generation of cross-lingual summaries using a novel method based on sentence selection, i.e. we present a novel compressive and extractive summarisation method. The technique is designed to reduce redundancy and improve the informativeness of cross-lingual summaries. An evaluation of the technique was found to outperform baseline results for the tested low-resourced languages. This work is currently under review and is presented in detail here.

In Section 3, a text summarisation technique based on neural cross-lingual abstractive methods is described. The solution makes use of a pretrained English model, which is fine-tuned to the target Slovene language.The technique was evaluated using both quantitative and human evaluation. Both the readability and accuracy of the generated summaries were found to be acceptable in this low-resourced language setting. The publication associated with this work is available in Appendix A.

In Section 4, our corpus summarisation visualisation *TeMoTopic* is described. This approach summarises corpus contents in terms of temporal topic slices and keywords. It enables interactive exploration of the associated documents with both the topics and keywords. This provides a summary overview of the corpus topics and the document details on demand. The publication associated with this work is provided in Appendix B.

We conclude the deliverable in Section 5. The final Section 6 contains the associated outputs of the work done within T4.2. All articles published as part of the task are enclosed in the appendices.

2 Extractive and Compressive Text Summarisation

This section focuses on extractive and compressive summarisation, as abstractive summarisation requires labeled corpora for training and understanding of natural language generation (McDonald, 2010; Mills & Bourbakis, 2013) and is covered in Section 3.

Nowadays, a wide variety of digital information from books, published articles, video or audio, is available online and keeps on increasing day by day (X. Li, Du, & Shen, 2012). For the lack of time, it becomes challenging for the readers to go through all such information to keep them up-to-date in a given topic of interest, which becomes even harder if the information is available only in unknown languages to the reader. To deal with these challenges, novel Cross Language Text Summarisation (CLTS) is required, which—from a given document or a set of documents in the source language—aims to generate a summary in a target language different from the source language (Zhu et al., 2019; J. Zhang et al., 2016).

Most of the existing CLTS systems (Wan, 2011; Jhaveri, Gupta, & Varma, 2019; Linhares Pontes et al., 2020; J. Zhang et al., 2016) have been proven powerful but failed to consider multiple perspectives together to improve the quality of the summary. For example, in a recent paper, Linhares Pontes et al. (2020) first select the sentences based on their highest score obtained using cross-language information access (called CoRank score (Wan, 2011)) and then replace them by their compression if available. Moreover, the abstractive CLTS systems (J. Zhang et al., 2016) use large annotated corpora for training, which is time-consuming. Here, the CLTS task is posed as a subset of the sentence selection problem. Recent studies show that for any subset selection problem, the multi-objective version outperforms the single objective version (Y. Zhang, Gong, Gao, Tian, & Sun, 2020; Oliveira, Sabourin, Bortolozzi, & Suen, 2002; Al-Tashi, Abdulkadir, Rais, Mirjalili, & Alhussian, 2020). Inspired by this, instead of single, multiple perspectives/objectives of the summary are simultaneously considered for their optimisation. For this purpose, we use the **M**ulti-objective **E**volutionary **A**lgorithm (Deb, 2015) (MEA) in a novel unsupervised way. Our unsupervised CLTS approach (named as **CLTS-MEA**) uses the binary differential evolution (DE) (Wang, Li, Li, & Wang, 2018) algorithm. While there exist other evolutionary algorithms like particle swarm optimisation (Du & Swamy, 2016), NSGA-II (Deb, Pratap, Agarwal, & Meyarivan, 2002), etc., DE has been highly successful[1] in various competitions organised under the IEEE Congress on Evolutionary Computation (CEC) conference series.

We pose CLTS as a binary optimisation problem, where the task is to select a subset of sentences from the target language such that they (a) leverage information (high co-rank score) from both the source and the target language; (b) exhibit high diversity among sentences to avoid redundancy in the summary. We also investigate a preference for longer sentences in the summary (Saini et al., 2019). At the end of the algorithm execution, a set of high-quality solutions (each solution representing a summary) are generated out of which the best solution (a summary) is selected. Another challenge in CLTS is to decide which perspectives will produce a good quality summary; therefore, we also study varying a combination of different perspectives.

As the arrangement of the sentences in the obtained summaries has a major role for their *readability*, the arrangement of the sentences in the final summary is analysed considering three scenarios in an unsupervised way. In the first one, sentences are arranged based on their position in the document or a sequence of documents, while in the second one, sentences are arranged based on their co-rank score (more details are provided in Section 2.4). In the third scenario, both position and co-rank score are considered to arrange the sentence. A readability score is calculated by observing the extent to which the next sentence $\mathscr{S}_t$ is related with the previous sentence $\mathscr{S}_{t-1}$ in the summary.

As reported in recent surveys (Pontes, Huet, Torres-Moreno, & Linhares, 2018; Linhares Pontes et al., 2020), researchers are also working on compressive approaches where sentences extracted based on their relevance are replaced by their compressions (removing non-relevant words) if they exist. Here, a compression can be either a single sentence or multi-sentences. In the first case, each sentence is analysed separately for the possible compression, while for multi-sentence compression, sentences are

[1] http://www.ntu.edu.sg/home/epnsugan/index_files/cec-benchmarking.htm

clustered based on their similarity ($> threshold$) in an incremental way, and then compressed for each cluster. Thus, multiple sentences (of a cluster) can have the same compression. In (Linhares Pontes et al., 2020), single and multi-sentence compression used the LSTM neural network (Hussain et al., 2019) and phrase-level chunking graph (Filippova, 2010) followed by the integer programming (Schrijver, 1998). It has also been shown that compressive approaches improve the summary evaluation score, grammatically and informativeness of the summary. To investigate the role of compressive approaches in our framework, we replaced the sentences of the best generated extractive summary with their multi-sentence compression (as it has been better than a single sentence compression in (Linhares Pontes et al., 2020)). Thus, we present the evaluation of extractive vs. compressive summaries and a use-case where we visualize the summary.

To evaluate the developed approach in terms of the standard ROUGE measure, as a source language we use six less-resourced European languages[2], including Finnish, Croatian, Estonian, Slovenian, Spanish, Portuguese, and one well/resourced language, French. In all the cases, the target language is English. In each language, there is a set of topics and each topic has a set of documents with three human written (also called gold) summaries. The results show that we are able to beat the state-of-the-art extractive and compressive CLTS systems.

2.1 Problem Definition

Let $\mathbb{T} = \{\mathfrak{D}_1, \mathfrak{D}_2, ...\mathfrak{D}_N\}$ be a topic consisting of a set of N documents in the target language. Here, $\mathfrak{D}_j$ is a j-th document and includes M sentences, $\{\mathcal{S}_{j,1}, \mathcal{S}_{j,2}, ..., \mathcal{S}_{j,M}\}$. We pose the cross-language (multi-document) summarisation task as a binary optimisation problem: The task is to select a subset of optimal L sentences $\{s_1, s_2, ..., s_L\}$, from a given set of documents $\mathbb{T}$, forming an extractive (compressive) summary $\mathcal{K}$ defined as

$$|\sum_{i=1}^{N}\sum_{j=1}^{M} x_{i,j}\mathcal{S}_{i,j}| \leq \mathfrak{L}_{max} \tag{1}$$

where $\mathfrak{L}_{max}$ is the maximum number of words allowed in the summary and $x_{i,j}$ indicate the presence or absence of j-th sentence of i-th documents in the summary, such that it maximises

$$maximise\left(\frac{1}{\Psi_1(\mathcal{K})}, \Psi_2(\mathcal{K}), \Psi_3(\mathcal{K})\right) \tag{2}$$

where, Ψ_1, Ψ_2, and Ψ_3 are different perspectives which are simultaneously optimised using the multi-objective optimisation (MOO) based binary differential evolution (MBDE) (for details, see Section 2.3). The mathematical definitions of these objective functions are as follows:
(a) Ψ_1 function is designed to maintain the diversity among the sentences in the summary and is expressed as

$$\Psi_1 = \left(\sum_{i=1}^{|\mathcal{K}|-1}\sum_{j=i+1}^{|\mathcal{K}|}\lambda(s_i, s_j)\right)/(|\mathcal{K}| - 1|) \tag{3}$$

where $|\mathcal{K}|$ is the total number of sentences in the summary, and $\lambda(s_i, s_j)$ is the similarity between the i-th and j-th sentence of the summary.
(b) Ψ_2**:** In the addressed cross-lingual setting, it is necessary to capture information from both the source and the target language. For this purpose, we used the CoRank (Wan, 2011) method, which calculates the relevance score of all sentences based on their similarity in each language individually and across both languages. Let $\mathbb{S}^{sr}$=$\{\mathscr{S}_1^{sr}, \mathscr{S}_2^{sr}, ..., \mathscr{S}_{\mathbb{S}}^{sr}\}$ be a set of merged source language sentences from all the documents in $\mathbb{T}$. Similarly, for the target language, $\mathbb{S}^{tr}$=$\{\mathscr{S}_1^{tr}, \mathscr{S}_2^{tr}, ..., \mathscr{S}_{\mathbb{S}}^{tr}\}$ is the set of sentence in the target language appearing in the same order as in the source set $\mathbb{S}^{sr}$. Thus, $|\mathbb{S}^{sr}|$=$|\mathbb{S}^{tr}|$= $|\mathbb{S}|$. Equations (4) to (8) are used to compute the ranking of the sentences in the source (target) using the target (source)

[2]https://cordis.europa.eu/project/id/825153

language. The salience scores u and v of the sentences belonging to the target and source languages, respectively, are computed as follows:

$$u = \beta.(\tilde{K^{tr}})^T u + \gamma.(\tilde{K}^{sr,tr})^T v \tag{4}$$

$$v = \beta.(\tilde{K^{sr}})^T v + \gamma.(\tilde{K}^{sr,tr})^T u \tag{5}$$

where β and γ are the relative contributions to the final salience score from the information in the same language and the information in the other language, such that $\beta + \gamma = 1$, and where K^{tr} and K^{sr} are the two affinity matrices showing the relationship between target language and source language sentences using the similarity scores shown below

$$K_{ij}^{tr} = \begin{cases} \lambda(\mathscr{S}_i^{tr}, \mathscr{S}_j^{tr}), & \text{if } i \neq j \\ 0 & otherwise \end{cases} \tag{6}$$

$$K_{ij}^{sr} = \begin{cases} \lambda(\mathscr{S}_i^{sr}, \mathscr{S}_j^{sr}), & \text{if } i \neq j \\ 0 & otherwise \end{cases} \tag{7}$$

$$K_{ij}^{sr,tr} = \sqrt{\lambda(\mathscr{S}_i^{tr}, \mathscr{S}_j^{tr}) \times \lambda(\mathscr{S}_i^{sr}, \mathscr{S}_j^{sr})} \tag{8}$$

In Equations 4 and 5, K^{tr} and K^{sr} are normalised to $\tilde{K}^{tr}$ and $\tilde{K}^{sr}$, respectively, to make the sum of each row equal to 1. The computation of Ψ_2 takes into account only the CoRank score of the target sentences as the output is in the target language.

$$\Psi_2 = \sum_{i=1}^{|\mathcal{K}|} u(\mathscr{S}_i) \quad \text{and} \quad \mathscr{S}_i \in \mathbb{S}^{tr} \tag{9}$$

(c) Ψ_3: The literature demonstrates the importance of longer sentences in the documents/blogs. We consider them in the proposed CLTS framework as

$$\Psi_3 = \sum_{i=1}^{|\mathcal{K}|} Length(\mathscr{S}_i) \quad \text{and} \quad \mathscr{S}_i \in \mathbb{S}^{tr}, \tag{10}$$

where $Length(\mathscr{S}_i)$ denotes the number of words in the i-th sentence of the summary after removing the stop words.

2.2 Datasets

For the evaluation, we used the English version of the MultiLing Pilot 2011 dataset (Giannakopoulos et al., 2011), which includes a range of topics and each topic is associated with a set of 10 English documents. Further, for every topic, three gold summaries, each of 250 words, are also provided in English. Previous studies (J.-g. Yao, Wan, & Xiao, 2015; Wan, 2011) showed that translating the source languages summaries generated by a monolingual summarisation method or summarising the document after translating the source documents is not a good idea. Therefore, this dataset is translated using the Google Translate service[3] into Finnish, Croatian, Estonian, Slovenian, Spanish, Portuguese, and French, which are used as test source languages in this work. Thereafter, the translations were manually checked to avoid any translation errors. The benefit of source and target languages are considered together in developing a cross-language multi-document summarisation system (L. Li & Li, 2013). Note that the CLTS task is different from the MultiLing Pilot 2011 task (where the source and target documents should be in the same language) organised by the Multilingual community[4].

[3] https://translate.google.com/
[4] http://multiling.iit.demokritos.gr

2.3 Methodology

We first describe the pre-processing of datasets, followed by the *CLTS-MEA* extractive summary generation. The methodology is illustrated using the pseudo code in Algorithm 1. Next, we describe the compressive summary generation from the extractive summary. The symbols used in the following sections are summarised in Table 1.

Table 1: Symbols with their descriptions.

Abbreviation	Description
$\mathbb{P}^t$	Population consisting of solutions at t-th generation
$\mathbb{Z}$	The number of solution in the population
$\mathbb{G}$	The maximum number of generations
$\|\mathbb{S}\|$	The total number of sentences
$\mathcal{K}$	The obtained summary
$\|\mathcal{K}\|$	The number of sentences in the summary
$\vec{\mathscr{S}}_i^{sr}/\vec{\mathscr{S}}_i^{tr}$	i-th Sentence vector in the source/target language
$\mid \mathbb{S}^{tr} \mid$	Total number of sentences in the target language
$\lambda(\mathscr{S}_i^{tr}, \mathscr{S}_j^{tr})$	Cosine similarity between two target's sentences
$\Gamma_{\|\mathbb{S}\|\times\|\mathbb{S}\|}$	Cosine distance matrix
$\mathbb{C}^{Rank}_{1\times\|\mathbb{S}\|}$	CoRank matrix of size $1\times \mid \mathbb{S} \mid$ using target sentences
$\mathfrak{L}_{max}$	Maximum number of words allowed in the summary

Pre-processing First, all the documents in a topic or event (considering the source and target language separately) are merged into a single document using their ordering in the datasets. The sentences from the source and target language are vectorised using the tf-idf scheme (W. Zhang, Yoshida, & Tang, 2011).

Construction of CoRank and similarity matrices. To assure the diversity of sentences in the summary, we used the following dissimilarity measure based on the cosine similarity:

$$\lambda(\mathscr{S}_i^{tr}, \mathscr{S}_j^{tr}) = \frac{\vec{\mathscr{S}}_i^{tr} . \vec{\mathscr{S}}_j^{tr}}{\| \vec{\mathscr{S}}_i^{tr} \| \| \vec{\mathscr{S}}_j^{tr} \|}, \tag{11}$$

where $\vec{\mathscr{S}}_i^{tr}$ and $\vec{\mathscr{S}}_j^{tr}$ are the (tf-idf) vectors of i-th and j-th sentences belonging to $\mathbb{S}^{tr}$. Here $\lambda(\cdot,\cdot)$ is the dot product of these two vectors. We compute the cosine similarity matrix $\lambda_{|\mathbb{S}|\times|\mathbb{S}|}$ and similarly, the CoRank matrix $\mathbb{C}^{Rank}_{1\times|\mathbb{S}|}$ as described in Section 2.1.

Sentence selection using CLTS-MEA. The pre-processed datasets and the generated matrices are utilised by the CLTS-MEA algorithm which can be divided into five parts briefly described below.

Solution representation and initialisation. Following the terminology used in DE, a chromosome/solution is represented as a binary vector where 1 at k-th position represents a presence of k-th sentence in the summary. DE starts from a set of randomly generated candidate solutions that form an initial population $\mathbb{P}$ of size $\mathbb{Z}$. The length of the solution is equal to the total number of sentences, i.e. $\mathfrak{L}_{max}$, Different perspectives (refer to Section 2.1) for each solution are also evaluated.

Genetic operators. Let us denote the current solution x_c and the population $\mathbb{P}$. We generate a set of offsprings ($\mathbb{Q}$) with the following constraint $|\mathbb{Q}| = 2 \times |\mathbb{P}|$. These is achieved through the methods current-to-best/1/bin and current-to-rand/1/bin, which are two methods among others in DE for generating offsprings; see (Mezura-Montes, Velázquez-Reyes, & Coello Coello, 2006; Wu et al., 2018). Specifically, the current-to-rand/1/bin helps in generating diverse solutions from the current solution, while current-to-best/1/bin provides a direction towards the currently best solution in the search space (Saini, Saha, Bhattacharyya, & Tuteja, 2020).

For both schemes, we generate probability vectors using the probability estimation operators (Wang et al., 2018). Then, the two generated probability vectors are converted into binary vectors (as we are in a binary space) $\mathbb{B}^t_{c,1}$ and $\mathbb{B}^t_{c,2}$ (called as trial vectors). The crossover operation is performed between the

Algorithm 1 CLTS-MEA Algorithm

Input: A collection of a set of documents in source and target language related to a specific event
Output: Summary in the target language

$\mathfrak{L}_{max}$ and $\mathbb{G} \leftarrow$ Maximum length of summary in terms of number of words and maximum number of generations
Compute $\lambda_{|\mathbb{S}|\times|\mathbb{S}|}$, $\mathbb{C}^{Rank}_{1\times|\mathbb{S}|}$, and $\mathbb{L}_{1\times|\mathbb{S}|}$ ▷ $|\mathbb{S}|$ is the total number of sentences after merging all the documents
for i←1 to $\mathbb{Z}$ **do** ▷ $\mathbb{Z}$ is the size of the population $\mathbb{P}$
    $\mathbb{P}[i] \leftarrow$ Initialise solution in binary space keeping $\mathfrak{L}_{max}$
    Compute Ψ_1, Ψ_2, and Ψ_3 for $\mathbb{P}[i]$
end for
$\mathbb{P}^1 \leftarrow \{\mathbb{P}[i], \mathbb{P}[2], \dots, \mathbb{P}[\mathbb{Z}]\}$ ▷ Initial population
for t←1 to $\mathbb{G}$ **do** ▷ $\mathbb{G}$ indicate the maximum number of generations
    $\mathbb{Q}^t \leftarrow \Phi$
    for c←1 to $\mathbb{Z}$ **do** ▷ c is the current solution number
        $\mathbb{Q}^t_{c,1}[i]$ and $\mathbb{Q}^t_{c,2}[i] \leftarrow$ Apply genetic operators on $\mathbb{P}^t[j]$ using current-to-rand/1/bin and current-to-best/1/bin scheme to give two new solutions
        Compute Ψ_1, Ψ_2, and Ψ_3 for $\mathbb{Q}^t_{c,1}$ and $\mathbb{Q}^t_{c,2}$
        Append $\mathbb{Q}^t_{c,1}$ and $\mathbb{Q}^t_{c,2}$ to $\mathbb{Q}^t$
    end for
    $\mathbb{R}_t \leftarrow \mathbb{Q}^t \cup \mathbb{P}^t$
    $\{F_1, F_2, \dots, F_M\} \leftarrow$ Apply non-dominated sorting on $\mathbb{R}^t$ to provide a set of non-dominated fronts
    $\mathbb{P}^{t+1} \leftarrow$ Select $\mathbb{Z}$ solutions considering rank-wise fronts and if needed, apply crowding distance operator
end for
Pick the solutions of the top front, i.e. $F1$
return The solution having the best extractive summary

current solution (x^t_c) and trial vector $\mathbb{B}^t_{c,1}$ ($\mathbb{B}^t_{c,2}$) to produce the new solution $\mathbb{Q}^t_{c,1}$ ($\mathbb{Q}^t_{c,2}$). For each current solution in the population, a new set of solutions is generated forming a new population $\mathbb{Q}^t$ consisting of $\{\mathbb{Q}^t_{1,1}, \mathbb{Q}^t_{1,2}, \mathbb{Q}^t_{2,1}, \mathbb{Q}^t_{2,2}, \dots, \mathbb{Q}^t_{\mathbb{Z},1}, \mathbb{Q}^t_{\mathbb{Z},2}\}$. Here, $\mathbb{Q}^t_{c,1}$ and $\mathbb{Q}^t_{c,2}$ are two new solutions corresponding to the current solution x^t_c at t-th generation and $c \in \{1, 2, \dots, \mathbb{Z}\}$.

Objective functions and environment selection. Our approach is based on the concept of multi-objective optimisation (MOO). Therefore, the quality measures (Ψ_1, Ψ_2, and Ψ_3) for the solutions in the new population $\mathbb{B}$ are calculated. We merge the parent and offspring population to form $\mathbb{R}^t = \mathbb{P}^t \cup \mathbb{Q}^t$. As a next step, non-dominated sorting (Deb et al., 2002) is performed on $\mathbb{R}$ to split it into non-dominated disjoint fronts -$\{F_1, F_2, \dots, F_M\}$ where $1 \leq M \leq |\mathbb{R}^t|$. The Front-1 (highest) includes the highest rank solutions and so on. Considering rank-wise fronts (highest to lower), $|\mathbb{P}|$ solutions are selected to form the population $\mathbb{P}$ for the next generation ($t+1$). In case of ties in selection of the solutions, we apply the crowding distance operator (Deb et al., 2002). We repeat the whole process of applying genetic operators and updating the population until we reach the maximum number of generations.

Summary selection. Our algorithm ends with a set of Pareto optimal solutions, all having equal importance. Any solution can be chosen based on the user's interest. In our case, we select the solution with the best (extractive) summary, which is further analysed for compression (discussed below). Later readability is studied for both types of summaries.

Compressive summary generation. Inspired by the results obtained by the multi-sentence compression (MSC) method for text summarisation, presented in deliverable D4.3 (Linhares Pontes et al., 2020) and in (Pontes et al., 2018; Linhares Pontes, Huet, Gouveia da Silva, Linhares, & Torres-Moreno, 2018), we investigate in this deliverable whether this method can improve the informativeness of already obtained summaries. MSC aims to generate a short sentence with the key information from a cluster of closely related sentences. In other words, MSC enables summarisation to generate outputs combining fully formed sentences from one or several documents. Below we briefly described the four major steps of the MSC used method.

Clustering. Clustering is the procedure of partitioning a set of objects into various groups based on a similarity/dissimilarity criterion. Here we consider an object as a sentence. In order to create clusters of similar sentences, we analyse the sentences in the source and target languages. Indeed, the repre-

sentation of sentences in multiple languages provides different analyses of their content, which enable us to obtain a better analysis of the similarity between the sentences (Giannakopoulos et al., 2011). Therefore, sentences are grouped in the same cluster if their similarity score is bigger than a threshold. The similarity score of a pair of sentences i and j is defined by the cosine similarity in both languages defined as:

$$sim(i,j) = \sqrt{\lambda(\mathscr{S}_i^{sr}, \mathscr{S}_j^{sr}) \times \lambda(\mathscr{S}_i^{tr}, \mathscr{S}_j^{tr})} \tag{12}$$

where $\mathscr{S}^{sr}$ and $\mathscr{S}^{tr}$ represent a sentence in the source and target languages, respectively.

Construction of Chunk-level Graph. Following the same idea as in (Linhares Pontes et al., 2020), we split the sentences at chunk level in order to keep the most useful structures. Then, we represent each cluster of similar sentences as a Chunk Graph (CG). Thereafter, we create the CG as described in (Linhares Pontes et al., 2020). Initially, this graph is composed of the first sentence, and the $-begin-$ and $-end-$ vertices. A chunk is represented by an existing vertex only if it has the same lowercase form, the same POS, and if there is no other chunk from that same sentence that has already been mapped onto that vertex. A new vertex is created if no vertex is found with its characteristics in the CG. Each sentence represents a simple path between the $-begin-$ and $-end-$ vertices. Sentences are analysed and added individually to the CG. For each analysed sentence, the chunks are inserted in the following order:

1. Chunks that are not stopwords and for which there is no unambiguous mapping candidate;
2. Chunks that are not stopwords and for which there are several possible candidates in the graph or that occur more than once in the same sentence;
3. Stopwords.

The arcs in the CG represent the cohesion between two chunks (Filippova, 2010). This cohesion is measured from the frequency and the position of these chunks in sentences:

$$w(i,j) = \frac{\text{cohesion}(i,j)}{\text{freq}(i) \times \text{freq}(j)}, \tag{13}$$

$$\text{cohesion}(i,j) = \frac{\text{freq}(i) + \text{freq}(j)}{\sum_{\mathscr{S} \in \mathbb{S}^{tr}} \text{diff}(\mathscr{S}, i, j)^{-1}}, \tag{14}$$

where $freq(i)$ is the chunk frequency mapped to the vertex i and the function $diff(\mathscr{S}, i, j)$ refers to the distance between the offset positions of chunks i and j in the sentence $\mathscr{S}$ of the cluster of similar sentences containing these two chunks. The higher the cohesion, the stronger is the relationship between the two chunks. An example of a created chunk graph using similar sentences is shown in Figure 1.

Finding Shortest Path. Filippova (2010) generated the compression of similar sentences by only calculating the shortest path in the graph. However, this procedure does not assure that the generated paths contain the main information of the CG. In order to generate the compression with the main information of these graphs, we consider keywords at the global (all documents) and local (cluster of similar sentences) levels to keep the main information of both the documents and the cluster of similar sentences. The keywords are selected by using LDA (Blei, Ng, & Jordan, 2003), a well-known topic based model that generates topics based on word frequency from a set of documents. We consider that each document belongs to only one topic and used the most relevant words that represent each topic.

Based on this analysis, we use the Integer Linear Programming (ILP) formulation described in (Linhares Pontes et al., 2018, 2020; Pontes, Huet, Torres-Moreno, da Silva, & Linhares, 2020) to find a path in CG that is composed of chunks with a good cohesion between them and with a maximum number of keywords. This is achieved with the following equation:

$$\text{Minimise} \Big(\sum_{(i,j) \in A} w(i,j) \cdot x_{i,j} - c \cdot \sum_{k \in K} b_k \Big) \tag{15}$$

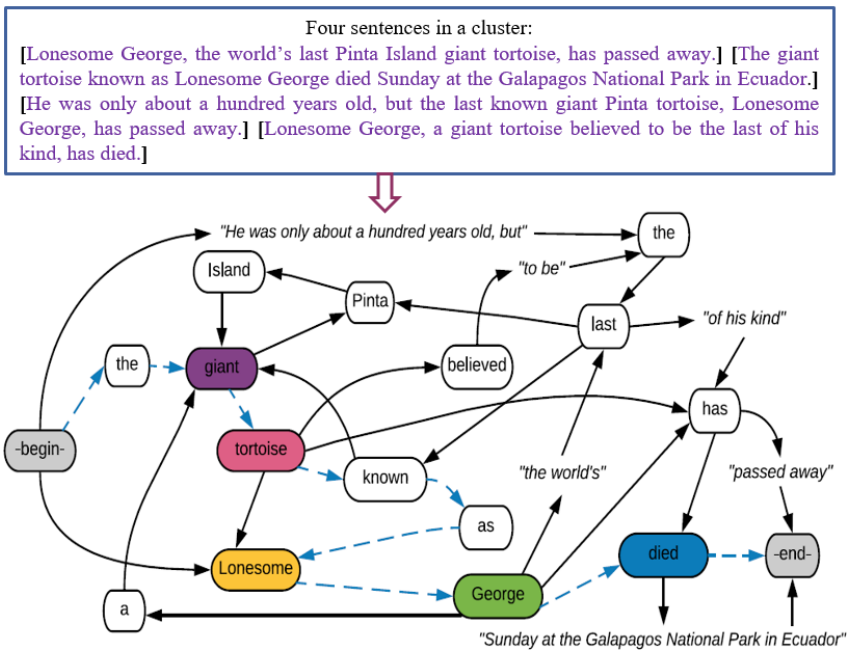


Figure 1: An example of creating a Chunk Graph extracted from (Linhares Pontes et al., 2020).

where x_{ij} indicates the existence of the arc (i, j) in the solution, $w(i, j)$ is the cohesion of the chunks i and j (Equation 13), K is the set of labels (each representing a keyword), b_k indicates the existence of a chunk containing the label (keyword) k in the solution and c is the keyword bonus of the graph[5].

We calculate the fifty best solutions according to the objective Equation 15 having at least eight chunks and at least one verb. The optimised score Equation 16 explicitly takes into account the size of the generated sentence. Finally, we select the sentence with the lowest final score obtained using Equation 16 as the best compression. For example, in Figure 1, the coloured dotted line is the identified as the shortest path

$$\text{score}_{norm}(\mathscr{S}) = \frac{e^{\text{score}_{opt}(\mathscr{S})}}{||\mathscr{S}||}, \tag{16}$$

where $\text{score}_{opt}(\mathscr{S})$ is the score of the sentence $\mathscr{S}$ from Equation 16.

Summary Generation. As an outcome of Section 2.3, we have the compression of each cluster, i.e. for each cluster, an informative sentence is generated by combining the information of the similar sentences. Thus, a set of similar sentences may have similar compression. To produce the final summary, we consider the sentences chosen by CLTS-MEA for building the extractive summary and we replace then with their compressed version. It should be indicated that not all the sentences may have a compression. For these cases the original version are used in the compressive summary.

2.4 Experimental Setup and Comparative Results.

This section starts by introducing the research questions that guided the experiments. This is followed by the evaluation measures, the parameters setting and the methods used for comparing the generation of extractive and compressive summaries. Finally, the experimental results are provided with their associated discussion.

Research questions. To guide the experiments, we list four research questions:

[5]The keyword bonus allows the generation of longer compression that may be more informative.

- *RQ1*: Are our obtained extractive and compressive summaries more informative than the existing baselines?
- *RQ2*: What will be the effect of using various combinations of the different perspectives and which one performs the best?
- *RQ3*: What about the readability of the obtained summaries in comparison with the summaries of the existing methods?
- *RQ4*: Which type of summary is better out of extractive and compressive?

Evaluation Measures. To check our generated summaries' informativeness, we have counted the common n-grams between our summaries and those in the gold summaries. In other words, we have used the REcall-Oriented Understudy for Gisting Evaluation (ROUGE) measure (Lin, 2004), which compares the distribution of words between the obtained summary and a set of available reference/gold summaries. For the value of 'n', we used 1-gram, 2-gram to provide ROUGE-1 and ROUGE-2, respectively. We also reported the ROUGE-L and ROUGE-SU4, which measures the Longest common sequence and skip units (SU), respectively. Note that the existing methods report only ROUGE-1, ROUGE-2, and ROUGE-SU4. The higher the value of these measures, the more informative is our summary. A summary $\mathcal{K}$ is readable if there is a smooth chain of sentences or in other words, the next sentence is related with the preceding sentence (Shareghi & Hassanabadi, 2008). Therefore, to measure it, we computed the readability factor ($\mathbb{RF}$) denoted as

$$\mathbb{RF}_{\mathcal{K}} = \sum_{0 \leq i < |\mathcal{K}|} Sim(\mathscr{S}_i, \mathscr{S}_{i+1}) \tag{17}$$

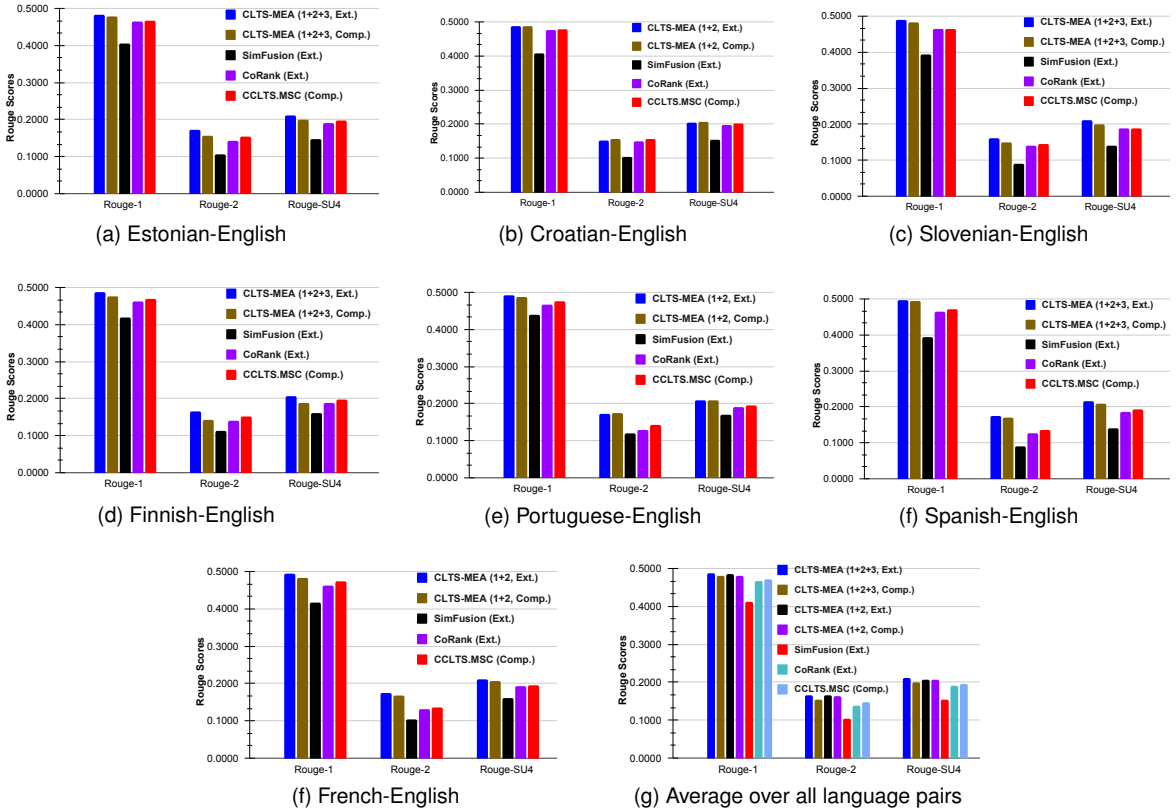
where, $Sim(.)$ is the similarity between preceding and next sentences in the semantic (embedding) space. To normalise it, we have divided $\mathbb{RF}_{\mathcal{K}}$ by the maximum similarity among the preceding and the next sentence (it varies for each language-pair).

Parameters Setting. We have executed our CLTS-MEA algorithm with the following parameter's value: the maximum number of generations is set to 50 and the number of solutions in the population is set to 25. These parameters are selected after a thorough sensitivity analysis. For the other parameters, F and CR, we have set a pool of values as $[0.6, 0.8, 1.0]$ and $[0.1, 0.2, 1.0]$, respectively. Both choices are motivated by Saini et al. (2020). Any value can be selected randomly for the generation of a new solution in each generation. To measure the readability of the summary in the semantic space, we have used the Universal Sentence Encoder[6] model to represent the sentences followed by cosine similarity calculation.

Comparative results on different Language Pair. For comparison, we have chosen two existing cross-language extractive summarisers, CoRank (Wan, 2011) and SimFusion (Wan, 2011). Also, we chose a recently developed compressive system, CCLTS.MSC (Linhares Pontes et al., 2020). A brief description of these methods is already provided at the beginning of Section **??**. Note that there are other compressive systems, such as (J.-g. Yao et al., 2015) and (Wan, Luo, Sun, Huang, & Yao, 2019), however CCLTS.MSC is one of the most recent ones. Moreover, the source code of (J.-g. Yao et al., 2015) and (Wan et al., 2019) is not available. Thus, it is hard to replicate their results as no public implementation is available; also these systems are language dependent. For extractive, we have chosen CoRank (Wan, 2011) and SimFusion (Wan, 2011) because they are language independent and have proven to be a strong baseline in the literature. The results of CCLTS.MSC are available only for some language pairs (Finnish, Croatian, Estonian, Slovenian, and French) but as the code is openly available[7], so we have re-run the approach for the remaining language pairs.

For the first *RQ1* and *RQ2*, we examine the performance of our proposed approach, CLTS-MEA, in terms of ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE SU4, over different language pairs (refer to Section 2.2). Table 2 lists the same for the extractive and compressive summaries. The ROUGE scores obtained after doing the ablation study using different perspectives are also listed in the table. Additionally,

[6] https://tfhub.dev/google/universal-sentence-encoder/1
[7] https://github.com/ElvysLPontes/Compressive-cross-language-text-summarisation



(a) Estonian-English (b) Croatian-English (c) Slovenian-English

(d) Finnish-English (e) Portuguese-English (f) Spanish-English

(f) French-English (g) Average over all language pairs

Figure 2: Comparison of our best results against the state-of-the-art systems on extractive and compressive summary generation. Seven language pairs are evaluated as well as their average. If our best results for a language pair is for the extractive strategy, then results of the corresponding compressive summary is also shown and vice-versa. The numbers 1, 2 and 3 denote the used objective functions Ψ_1, Ψ_2 and Ψ_3, respectively.

a detailed comparison with the existing extractive and compressive systems is shown in Figure 2. It is worth noticing that for Estonian-English, Spanish-English, Finnish-English, and Slovenian-English, our extractive approach utilising three perspectives (Ψ_1, Ψ_2, and Ψ_3) together outperforms the existing alternatives. While for French-English and Portuguese-English language pairs, our extractive system obtains the top score but utilising only two perspectives (Ψ_1, Ψ_2). Only for the Croatian-English pair, our compressive summary generation method shown to have a top performance, where the opted perspectives were Ψ_1 and Ψ_2. In other words, the combination of Ψ_1 and Ψ_3 contribute the less in the summary performance but when considering Ψ_2 along with them, there is a gain in terms of ROUGE-1, ROUGE-2, and ROUGE-SU4 scores over all language pairs. The best ROUGE scores for each language pair are highlighted in bold in Table 2.

In terms of relative improvement (%), our best result in Table 2 outperforms the recent CCLTS.MSC method in terms of ROUGE-1, ROUGE-2, and ROUGE-SU4, as follows: (a) Estonian-English: (3.19, 12.4, 7.51); (b) Croatian-English: (1.89, 1.43, 3.08); (c) Slovenian-English: (5.26, 11.5, 11.5); (d) Finnish-English: (3.84, 8.80, 5.56); (e) Portuguese-English: (3.33, 22.02, 7.23); (f) Spanish-English: (5.19, 28.5, 12.23); (g) French-English: (3.98, 27.64, 11.05). It is worth noting that for some of the language pairs, the best ROUGE-1, ROUGE-2, and ROUGE-SU4 scores are corresponding to different methods or combinations of the perspectives. For example, (a) for Croatian-English pair, ROUGE-1 and ROUGE SU4 are better corresponding to CLTS-MEA(Ψ_1,Ψ_2), but ROUGE-2 is better corresponding to CLTS-MEA(Ψ_1,Ψ_2,Ψ_3); (b) for French-English pair, ROUGE-1 and ROUGE-2 are better using CLTS-MEA (Ψ_1,Ψ_2), but ROUGE-SU4 is better using CLTS-MEA(Ψ_1,Ψ_2,Ψ_3). In this case, it becomes challenging to decide which system is best. To answer this question, we have averaged the ROUGE scores cor-

Table 2: The results obtained with the proposed approach for different language pairs. Ext. and Comp. stands for extractive and compressive summaries, *R* stands for Recall, and *Obj*. indicates the number of objectives used for optimisation.

Language Pair→		Estonian-English				Croatian-English				Slovenian-Eng			
Obj.↓	**Ext./Comp.↓**	**R-1**	**R-2**	**R-L**	**R-SU4**	**R-1**	**R-2**	**R-L**	**R-SU4**	**R-1**	**R-2**	**R-L**	**R-SU4**
Ψ_1, Ψ_2, Ψ_3	**Ext.**	**0.4825**	**0.1730**	0.4491	**0.2120**	0.4867	**0.1589**	0.4549	0.2034	**0.4901**	0.1618	**0.4640**	**0.2109**
	Comp.	0.4775	0.1573	0.4450	0.1994	0.4811	0.1585	0.4522	0.2043	0.4819	0.1492	0.4516	0.1994
Ψ_1, Ψ_2	**Ext.**	0.4824	0.1698	**0.4580**	0.2099	0.4862	0.1510	0.4544	0.2048	0.4767	**0.1628**	0.4511	0.2005
	Comp.	0.4732	0.1718	0.4440	0.2089	**0.4882**	0.1571	**0.4587**	**0.2072**	0.4768	0.1617	0.4499	0.2011
Ψ_1, Ψ_3	**Ext.**	0.4569	0.1481	0.4264	0.1884	0.4495	0.1318	0.4181	0.1797	0.4421	0.1436	0.4148	0.1849
	Comp.	0.4347	0.1213	0.4022	0.1629	0.4313	0.1155	0.3985	0.1650	0.4263	0.1308	0.3922	0.1677

Language Pair→		Finnish-English				Portuguese-English				Spanish-English			
Obj.↓	**Ext./Comp.↓**	**R-1**	**R-2**	**R-L**	**R-SU4**	**R-1**	**R-2**	**R-L**	**R-SU4**	**R-1**	**R-2**	**R-L**	**R-SU4**
Ψ_1, Ψ_2, Ψ_3	**Ext.**	**0.4862**	**0.1645**	**0.4577**	**0.2074**	0.4854	0.1574	0.4531	0.2079	**0.4964**	**0.1752**	**0.4644**	**0.2156**
	Comp.	0.4758	0.1416	0.4405	0.1887	0.4767	0.1501	0.4470	0.2005	0.4950	0.1698	0.4602	0.2091
Ψ_1, Ψ_2	**Ext.**	0.4824	0.1609	0.4497	0.2066	**0.4911**	**0.1746**	**0.4651**	**0.2099**	0.4895	0.1619	0.4559	0.2073
	Comp.	0.4765	0.1597	0.4478	0.2029	0.4882	0.1742	0.4607	0.2097	0.4843	0.1585	0.4512	0.2035
Ψ_1, Ψ_3	**Ext.**	0.4699	0.1616	0.4390	0.2021	0.4447	0.1242	0.4137	0.1713	0.4447	0.1452	0.4103	0.1817
	Comp.	0.4511	0.1482	0.4189	0.1851	0.4190	0.1078	0.3862	0.1544	0.4207	0.1227	0.3862	0.1651

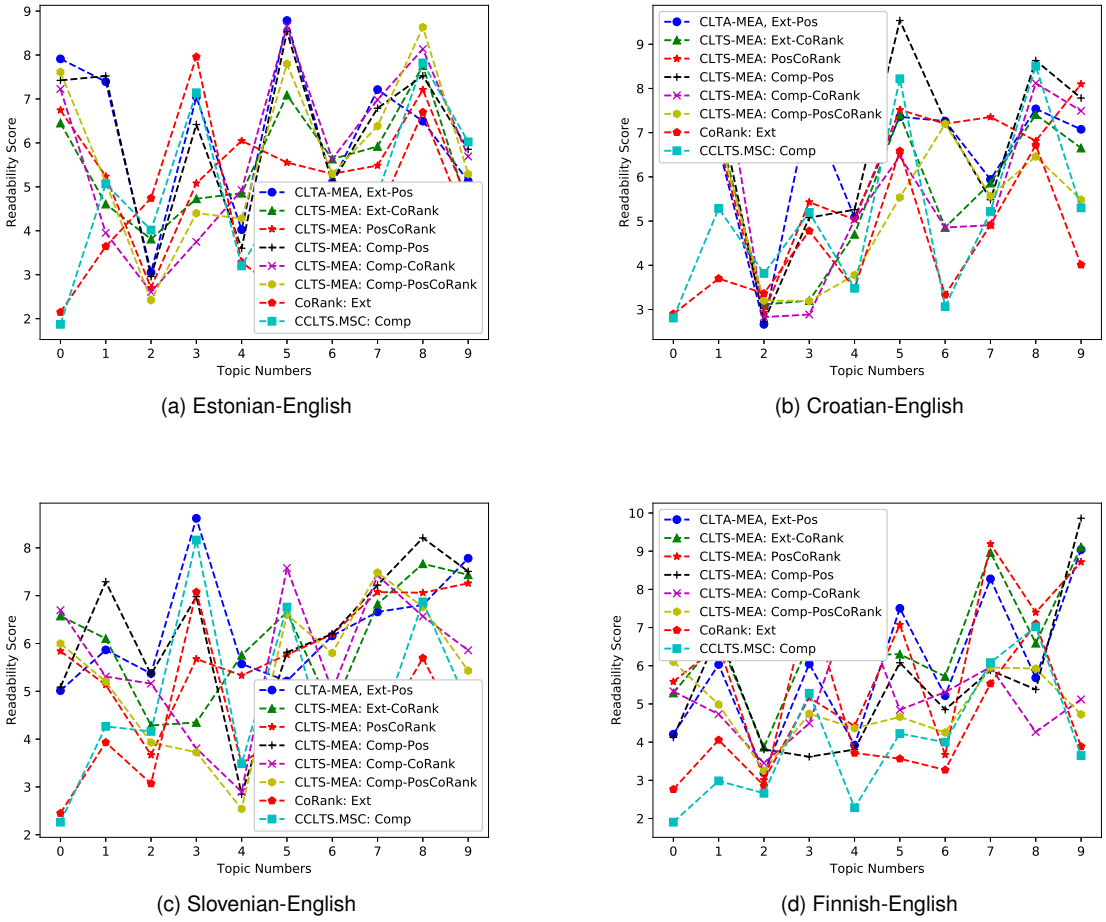
Language Pair→		French-English			
Obj.↓	**Ext./Comp.↓**	**R-1**	**R-2**	**R-L**	**R-SU4**
Ψ_1, Ψ_2, Ψ_3	**Ext.**	0.4913	0.1715	0.4561	**0.2162**
	Comp.	0.4799	0.1562	0.4462	0.2002
Ψ_1, Ψ_2	**Ext.**	**0.4932**	**0.1747**	**0.4664**	0.2120
	Comp.	0.4828	0.1676	0.4573	0.2077
Ψ_1, Ψ_3	**Ext.**	0.4616	0.1370	0.4223	0.1821
	Comp.	0.4307	0.1207	0.3958	0.1657

responding to different methods for extractive and compressive summaries and then, performed the ranking of the systems considering each ROUGE (1, 2, and SU4) measure individually in Table 3. For example, in column 3 of Table 3, different systems are ranked based on the highest to lowest ROUGE-1; the ranks are written in parentheses. Then, we considered the average of ranks per method as a way to identify a robust summariser. For example, 1, 3, 5 are the rank of a method as per ROUGE-1, 2, and SU4, respectively, then rank will be $(1 + 3 + 5)/3$. The first and second rank was achieved by the CLTS-MEA (Ψ_1,Ψ_2,Ψ_3) and CLTS-MEA (Ψ_1,Ψ_2) generating extractive summary. The method CCLTS.MSC has obtained a 5-th rank. Thus, the highest ROUGE scores (0.4884, 0.1665, and 0.2105 for ROUGE-1, 2, and SU4, respectively) and top ranking (1.33) of our system utilising Ψ_1,Ψ_2, and Ψ_3 proves that our extractive summary is more informative than the existing ones which are the answers to our research questions *RQ*1 and *RQ*2.

Table 3: Ranking of different systems including ours. Ranks w.r.t. other summarisers are presented in parentheses. 1 represents the highest rank and so on.

Method	Ext./Comp.	ROUGE-1	ROUGE-2	ROUGE-SU4	Rank
CLTS-MEA (Ψ_1, Ψ_2, Ψ_3)	Ext.	0.4884 (1)	0.1660 (2)	0.2105 (1)	1.33
	Comp.	0.4811 (4)	0.1547 (5)	0.2002 (4)	4.33
CLTS-MEA (Ψ_1, Ψ_2)	Ext.	0.4859 (2)	0.1646 (3)	0.2073 (2)	2.33
	Comp.	0.4814 (3)	0.1643 (4)	0.2059 (3)	3.33
CLTS-MEA (Ψ_1, Ψ_3)	Ext.	0.4528 (7)	0.1417 (7)	0.1843 (8)	7.33
	Comp.	0.4305 (8)	0.1665 (1)	0.1981 (5)	4.67
SimFusion	Ext.	0.4110 (9)	0.1038 (8)	0.1533 (9)	8.67
CoRank	Ext.	0.4659 (6)	0.1376 (9)	0.1905 (7)	7.33
CCLTS.MSC	Comp.	0.4717 (5)	0.1463 (6)	0.1952 (6)	5.67

A Study on Readability of the Summaries. To answer *RQ*3, we have plotted the average readability factor (Equation 17) for every topic in each language pair used as illustrated in Figures 3, 4, and 5. Note that (a) we have considered CLTS-MEA, optimising Ψ_1, Ψ_2, and Ψ_3 as this combination gains the first rank considering average ROUGE score over all language pairs (Table 3); (b) both extractive and compressive generated summaries are considered. For both types of summaries, we have arranged the sentences of the summaries based on their (a) position (*Pos*) in the merged documents having the con-



(a) Estonian-English

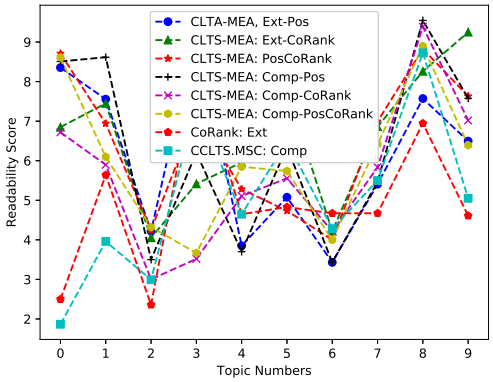
(b) Croatian-English

(c) Slovenian-English

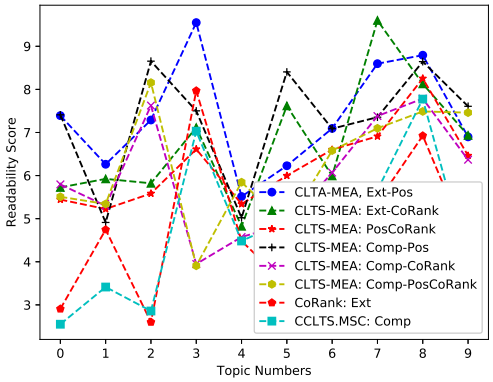
(d) Finnish-English

Figure 3: Readability factor over different topics of the language pairs (Estonian-English, Croatian-English, Slovenian-English, and Finnish-English) and average readability factor/score over all topics corresponding to summaries of different methods. Here, CLTS-MEA: Ext Pos/CoRank/PosCoRank indicate the obtained extractive summary by CLTS-MEA having arrangement of sentences based on position, co-rank, and position+corank score, respectively. Similarly for compressive (Comp) summary.

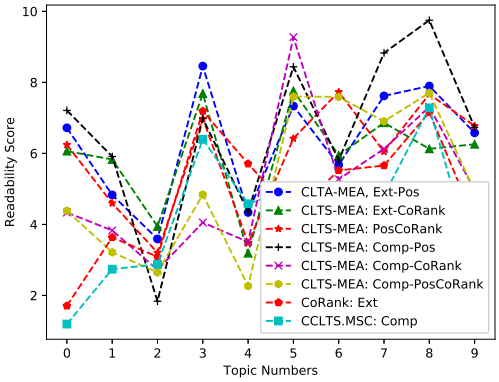
tent of all the documents under a topic/event; (b) *corank* score of the sentences; and, (c) based on the combined score of position and co-rank score, to investigate the effect. As each one may performs good for a range of topics and it's difficult to say exactly which one is good; therefore, we have taken the average of $\mathbb{RF}$ over all methods for a language pair and shown in Figure 2(g). From this figure, following things are inferred in terms of better method, summary (out of extractive and compressive) followed by order of arrangement of sentences and readability scores: (a) Croatian-English: CLTS-MEA, Compressive summary, *Pos*, 6.3728; (b) Estonian-English: CLTS-MEA, Compressive summary, *corank*, 5.7055; (c) Slovenian-English: CLST-MEA, extractive summary, *Pos*, 5.7843; (d) Finnish-English: CLST-MEA, extractive summary, *Pos*, 5.7930; (e) Portuguese-English: CLST-MEA, Compressive summary, *Pos+corank*, 6.3355; Spanish-English: CLST-MEA, Compressive summary, *Pos*, 7.3635; and (f) French-English: CLST-MEA, Compressive summary, *Pos*, 5.7930. Undoubtedly, in most cases, the compressive summary is better in terms of readability. This effect is follow the motivation to use compressive methods as they remove the irrelevant words of the final sentences. Moreover, our approach outperforms the compressive alternatives (CCLTS.MSC) (Linhares Pontes et al., 2020) and extractive CoRank (Wan, 2011) approaches by obtaining a readability score of (a) $(5.8438, 4.7971)$, (b) $(4.0805, 4.1336)$, (c) $(4.0460, 3.6644)$, (d) $(3.5723, 3.9415)$, (e) $(4.5034, 4.2234)$, (f) $(4.5690, 4.5265)$, (g) $(3.9564, 4.6370)$, for Croatian-English, Estonian-English, Slovenian-English, Finnish-English, Portuguese-English, Spanish-English, and French-English



(e) Portuguese-English



(f) Spanish-English



(f) French-English

Figure 4: Readability factor over different topics of the language pairs (Portuguese-English, Spanish-English, and French-English) and average readability factor/score over all topics corresponding to summaries of different methods. Here, CLTS-MEA: Ext Pos/CoRank/PosCoRank indicate the obtained extractive summary by CLTS-MEA having arrangement of sentences based on position, co-rank, and position+corank score, respectively. Similarly for compressive (Comp) summary.

language pairs, respectively. Moreover, arranging sentences based on their position is found to have a positive impact in most of the cases; therefore, results reported in Table 2 are the same.

Analysis on our Extractive and Compressive Summary with an Use Case. Now, we turn to our *RQ*4 using an example analysis. In Figure 6, we show an example of extractive and compressive summaries obtained using our CLTS-MEA (Ψ_1, Ψ_2, Ψ_3) corresponding to the Topic-3 of the Spanish-English languages pair. The corresponding actual summaries (three reference summaries) are also in part 'a' of the same figure. For the extractive summary (based on the position, as it has the good readability discussed in the above section and also can be analysed by reading the summaries) shown in the figure, we have obtained the average ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 as 0.4963, 0.1752, 0.4644, and, 0.2156, respectively. While for the observed compressive summaries, the same measures are as 0.4950, 0.1698, 0.4602, and 0.2091, respectively. We can infer that both the extractive and compressive summaries have a highly similar Rouge-1 and Rouge-L scores. However, in terms of Rouge-2 and Rouge-SU4, the extractive summaries have a gain of 3.18% and 3.5% with respect to the compressive ones. For comparative analysis, the summaries generated by CoRank and CCLTS.MSC methods of the same topic and language pair, are also shown in Figure 7 which have lower ROUGE scores than our method.

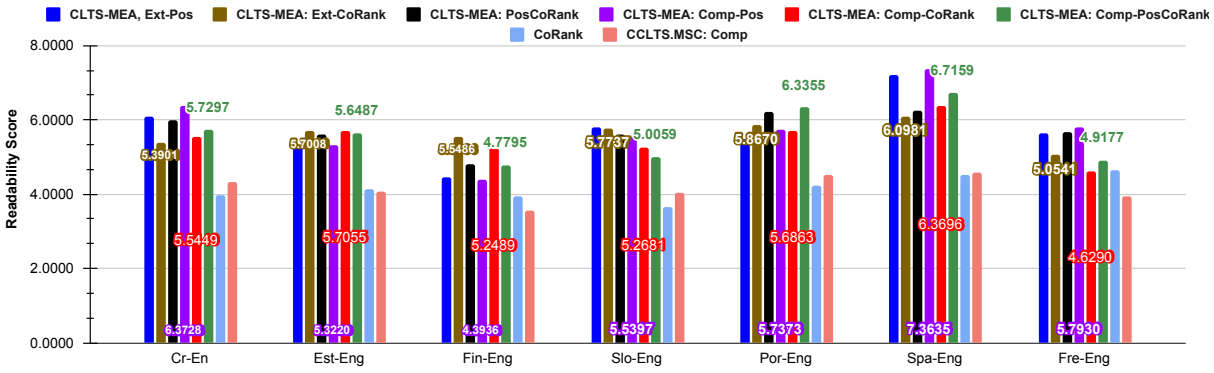


Figure 5: Average over different topics of language pairs. Readability factor over different topics of the language pairs and average readability factor/score over all topics corresponding to summaries of different methods. The numbers in represent the maximum value of bar of same colour as of number.

This proves that for the used dataset, the extractive summaries perform better than the compressive summaries.

Statistical Significance t-test. To check the superiority of our algorithm, *CLTS-MEA*, corresponding to the best results of each dataset over the other methods, we have performed the statistical significance test. There exist many tools to measure this like ANOVA (Mishra, Singh, Pandey, Mishra, & Pandey, 2019), paired t-test (Chan, Cheng, Mead, & Panjer, 1973), among others. We have chosen the t-test at 5% significance level which considers two groups. It includes two hypotheses: *null* and *alternative*. The first one considers that there are insignificant differences between the mean values of two groups, while the later one, says the reverse. As an outcome, it provides p-values and a lower value ($p < 0.05$) signifies the rejection of null hypothesis or in other words, it is the indication of superiority of our algorithm.

We have considered two groups: (a) a set of values by our best method, i.e. the ROUGE (1/2/SU4) scores corresponding to the CLTS-MEA optimising (Ψ_1, Ψ_2, Ψ_3) for extractive summary generation and (b) a set of ROUGE (1/2/SU4) values of the existing method. The obtained p-values are shown in Table 4. From this table, it is clear that for all the data sets, p-values are smaller than 0.05 which indicate to reject the null hypothesis and thus demonstrate the potentiality of our algorithm CLTS-MEA following the evolutionary procedure. Only for Croatian-English pair, the p-values corresponding to ROUGE-2 and ROUGE-SU4 are not significant as the best results are nearly the same as the existing methods. However, in terms of ROUGE-1, there is a significant improvement.

Complexity Analysis. We have analysed the worst-case complexity of our approach, CLTS-MEA. Below, we will discuss the complexity of the steps-3 to 16 of Algorithm 1: **(a)** as population initialisation (step-3 to 5) takes place in a binary space, so it takes $\mathcal{O}(\mathbb{Z})$ time, and corresponding calculation of different M perspectives for each solution takes $\mathcal{O}(\mathbb{Z}M)$. Hence, the time complexity of population initialisation is $\mathcal{O}(\mathbb{Z} + \mathbb{Z}\mathbb{M})$ which is equivalent to $\mathcal{O}(\mathbb{Z}M)$; **(b)** *Offspring* generation (steps-9 to 12) using two different schemes of DE takes $\mathcal{O}(\mathbb{Z}) + \mathcal{O}(\mathbb{Z})$ ignoring the time required arithmetic operations. Further, each offspring by objective function calculation which takes $\mathcal{O}(2\mathbb{Z}M)$ time; **(c)** merging the old and *Offspring*

Table 4: The p-values obtained by comparing the best results (ROUGE Scores reported in Table 2) of each language pair with the existing methods.

	p-values								
Ist Group+ IInd Group→	**CLTS-MEA + SimFusion**			**CLTS-MEA + CoRank**			**CLTS-MEA + CCLTS.MSC**		
Language-pairs↓	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
Estonian-English	5.68E-102	9.08E-80	4.51E-74	2.68E-08	5.11E-19	2.07E-10	1.14E-05	1.58E-08	1.31E-05
Croatian-English	3.75E-106	2.3E-55	2.3E-55	0.00034	0.0195	0.0666	0.00973	*0.673*	*0.0934*
Slovenian-English	4.64E-143	1.19E-91	8.46E-89	2.4E-14	2.53E-10	4.12E-12	2.37E-13	6.47E-07	7.61E-11
Finnish-English	8.86E-79	1.57E-52	7.62E-44	7.23E-13	2.97E-13	3.68E-09	9.00E-08	9.88E-05	0.00165
Portuguese-English	2.63E-51	1.01E-55	7.33E-31	7.23E-13	3.09E-41	4.53E-08	3.04E-06	6.49E-21	3.45E-05
Spanish-English	1.26E-153	3.42E-118	2.42E-95	1.49E-20	1.67E-45	4.22E-18	2.37E-13	7.33E-31	2.16E-12
French-English	5.01E-98	2.15E-88	3.01E-58	2.59E-20	2.33E-36	1.39E-10	1.89E-08	2.68E-29	1.39E-10

Reference Summary1: [Alberto Contador of Spain won the 2007 Tour de France general classification.] [He won by 23 seconds over the second place finisher Cadel Evans and by 31 seconds over Levi Leipheimer, who finished in the third place.] [Mauricio Soler won the King of the Mountains classification.] [Discovery Channel Pro Cycling Team won the team competition, while Caisse d'Epargne and Team CSC finished second and third, respectively.] [Le Grand Depart of the 2007 Tour de France was held in London, where the Prologue and Stage One of the Tour had been hosted on 7th and 8th of July, respectively.] [The sun and the crowds came out to welcome the Tour to London at both days.] [However, some incidents overshadowed the event.] [On 25 July, two explosive devices were found on a section of the Tour de France bicycle race just outside the Spanish town of Belagua.] [The Basque ETA claimed responsibility for placing the devices on the route.] [At the same day, Michael Rasmussen had been withdrawn from the Tour by his team Rabobank -although he had been tested negative for doping - for violating the team's internal rules.] [Rasmussen, who kept his overall lead in stage 15 by 2' 23' over the final winner Contador, had lied to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [Previously, Astana Team had announced that Alexander Vinokourov failed a blood test.] [Vinokourov, who won stages 13 and 15, was a pre-Tour favorite.]

Reference Summary2: [On July 7-8, 2007 Tour de France took place in London, the parts of the Tour that London will be hosting are the Prologue and Stage One.] [The starting points for these two events were the Trafalgar Square and Greenwich respectively.] [During the Tour de France Prologue half a million onlookers gathered and 180 plus riders participated.] [The winner was Cancellara completing 7.9 kilometer in 8`50``.] [The first round was completed with winners McEwan and Cancellara.] [Many people attended the event taking advantage of the good weather.] [At Jubilee Gardens the atmosphere was festive, the policing was low key, with many events and participants from France and England.] [In the third and longest stage of the Tour de France Cancellara maintained the first place.] [Vinokourov the winner of 13 and 15 stages failed the blood test, as they detected a double population of haematids, which implies there has been a blood transfusion, for that they have asked Astana, Vinokourov's Team to leave the race.] [Also, Rasmussen was suspended by the Danish Cycling Federation, as he refused to pass anti-doping control.] [Rasmussen said that he was in Mexico from 4 till 26 of June, while Cassani, a TV commentator, saw him in a training camp in Italy on June 13/14.] [He withdrew from the organization.] [On July 25 2007 explosive devices were found on a section of the Tour de France bicycle race near the town of Belagua.] [Alberto Contador was the winner of the Tour de France.]

Reference Summary3: [Le Grand Depart of the 2007 Tour de France was held in London, where the Prologue and Stage One of the Tour had been hosted on 7th and 8th of July, respectively.] [During this institution, Cancellara won twice in time trial and had been acclaimed by the audience.] [Despite the shadow of terrorist attacks, policing was discreet.] [In the event, Michael Rasmussen had been withdrawn from the Tour by his team Rabobank as he had violated internal regulations of his group.] [Rasmussen, who kept his overall lead in stage 15 by 2' 23' over the final winner Contador, had lied to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [In cycling team, Discovery Channel won.] [Hincapie led his team and led it to victory.] [Astana cycling team announced that Vinokourov failed a blood test following his victory in time trial.] [Tour organizers had asked Astana Cycling Team to leave the race, which had been accepted spontaneously.] [Vinokourov denied any blood manipulation.] [The event caused a lot of comments.] [Two explosive devices were found on a section of the Tour de France bicycle race just outside the Spanish town of Belagua.] [There were no reports of injuries but conflicting reports mentioned that the cyclists had passed the area before the devices went off, while others reported that the devices were detonated before the cyclists passed.] [ETA claimed responsibility for placing the devices on the route.]

(a) Actual/Reference Summaries in English.

Based on Position: [Alexander Vinokourov of Kazakhstan has won stage 13 of the 2007 Tour de France.] [Michael Rasmussen of Denmark held on to his overall lead, finishing 11th in the stage.] [This stage provided the first shake-up in the standings in days.] [Tomorrows stage will take the race into the Pyrenees.] [Tour de France: Yellow jersey Cancellara surprises sprinters.] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [Alberto Contador of Spain has won the 2007 Tour de France general classification.] [Alexandre Vinokourov of Kazakhstan has won stage 15 of the 2007 Tour de France in a time of 5h 34 28.] [ETA places explosives on Tour de France route in Spain.] [For a quarter of an hour afterwards tour buses, and still yet more support vehicles followed in the wake of the cyclists.] [The route, which will run through the Greenwich Millennium Village, is being prepared for the Départ.] [The parts of the Tour that London will be hosting are the Prologue and Stage One.] [Stage One will be on the following day, starting in Greenwich at 11:00 BST and finishing in Canterbury, Kent.] [At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [Coming to England for the first time since 1994.]

Based on CoRank:
[At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [Alexander Vinokourov of Kazakhstan has won stage 13 of the 2007 Tour de France.] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [Alexandre Vinokourov of Kazakhstan has won stage 15 of the 2007 Tour de France in a time of 5h 34 28.] [The route, which will run through the Greenwich Millennium Village, is being prepared for the Départ.] [Alberto Contador of Spain has won the 2007 Tour de France general classification.] [The parts of the Tour that London will be hosting are the Prologue and Stage One.] [Stage One will be on the following day, starting in Greenwich at 11:00 BST and finishing in Canterbury, Kent.] [For a quarter of an hour afterwards tour buses, and still yet more support vehicles followed in the wake of the cyclists.] [Michael Rasmussen of Denmark held on to his overall lead, finishing 11th in the stage.] [ETA places explosives on Tour de France route in Spain.] [Coming to England for the first time since 1994.] [Tour de France: Yellow jersey Cancellara surprises sprinters.] [This stage provided the first shake-up in the standings in days.] [Tomorrows stage will take the race into the Pyrenees.]

Based on Postion+CoRank:
[At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [The route, which will run through the Greenwich Millennium Village, is being prepared for the Départ.] [The parts of the Tour that London will be hosting are the Prologue and Stage One.] [Stage One will be on the following day, starting in Greenwich at 11:00 BST and finishing in Canterbury, Kent.]] [Alexandre Vinokourov of Kazakhstan has won stage 15 of the 2007 Tour de France in a time of 5h 34 28.] [Coming to England for the first time since 1994.] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [For a quarter of an hour afterwards tour buses, and still yet more support vehicles followed in the wake of the cyclists.] [Alexander Vinokourov of Kazakhstan has won stage 13 of the 2007 Tour de France.] [Alberto Contador of Spain has won the 2007 Tour de France general classification.] [ETA places explosives on Tour de France route in Spain.] [Tour de France: Yellow jersey Cancellara surprises sprinters.] [Michael Rasmussen of Denmark held on to his overall lead, finishing 11th in the stage.] [Tomorrows stage will take the race into the Pyrenees.] [This stage provided the first shake-up in the standings in days.]

(b) Generated Extractive Summaries based on different features.

Based on Position: [Alexandre vinokourov of kazakhstan has won the 2007 tour de france.] [Michael Rasmussen of Denmark held on to his overall lead, finishing 11th in the stage.] [This stage provided the first shake-up in the standings in days.] [Tomorrows stage will take the race into the Pyrenees.] [tour de france : yellow jersey rasmussen withdrawn .] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [alexandre vinokourov of kazakhstan has won the 2007 tour de france .] [ETA places explosives on Tour de France route in Spain.] [For a quarter of an hour afterwards tour buses, and still yet more support vehicles followed in the wake of the cyclists.] [At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [The Tour had to compete with more familiar sporting events, the British Grand Prix, the closing stages of Wimbledon and with the Live Earth concerts for the hearts of the British public.] [Run on the second anniversary of the 2005 terrorist attacks and at a time of heightened security, policing was successfully discrete the most visible police presence, by way of their novelty being, 45 members of the Gendarme Nationale.] [london and the south east for the two days , the sun and the crowds came out to welcome the tour de france to london.] [Londoners may get to see todays riders on their two wheels, but they will be followed by 1,500 vehicles, 13,000 policemen and women patrolling the route and 2,300 members of the world press.]

Based on CoRank: [Alexandre vinokourov of kazakhstan has won the 2007 tour de france.] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [tour de france : yellow jersey rasmussen withdrawn.] [london and the south east for the two days , the sun and the crowds came out to welcome the tour de france to london.] [At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [The Tour had to compete with more familiar sporting events, the British Grand Prix, the closing stages of Wimbledon and with the Live Earth concerts for the hearts of the British public.] [Run on the second anniversary of the 2005 terrorist attacks and at a time of heightened security, policing was successfully discrete the most visible police presence, by way of their novelty being, 45 members of the Gendarme Nationale.] [For a quarter of an hour afterwards tour buses, and still yet more support vehicles followed in the wake of the cyclists.] [Michael Rasmussen of Denmark held on to his overall lead, finishing 11th in the stage.] [ETA places explosives on Tour de France route in Spain.] [Londoners may get to see todays riders on their two wheels, but they will be followed by 1,500 vehicles, 13,000 policemen and women patrolling the route and 2,300 members of the world press.] [This stage provided the first shake-up in the standings in days.] [Tomorrows stage will take the race into the Pyrenees.]

Based on Position+CoRank: [At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [london and the south east for the two days , the sun and the crowds came out to welcome the tour de france to london.] [The Tour had to compete with more familiar sporting events, the British Grand Prix, the closing stages of Wimbledon and with the Live Earth concerts for the hearts of the British public.] [Run on the second anniversary of the 2005 terrorist attacks and at a time of heightened security, policing was successfully discrete the most visible police presence, by way of their novelty being, 45 members of the Gendarme Nationale.] [Londoners may get to see todays riders on their two wheels, but they will be followed by 1,500 vehicles, 13,000 policemen and women patrolling the route and 2,300 members of the world press.] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [For a quarter of an hour afterwards tour buses, and still yet more support vehicles followed in the wake of the cyclists.] [alexandre vinokourov of kazakhstan has won the 2007 tour de france.] [ETA places explosives on Tour de France route in Spain.] [tour de france : yellow jersey rasmussen withdrawn.] [Michael Rasmussen of Denmark held on to his overall lead, finishing 11th in the stage.] [Tomorrows stage will take the race into the Pyrenees.] [This stage provided the first shake-up in the standings in days.]

(c) Generated Compressive using Extractive Summaries based on different features.

Figure 6: A use case showing reference summaries along with extractive and compressive summary generated by CLTS-MEA (Ψ_1, Ψ_2, Ψ_3). Here, '[]' represents the sentence boundaries.

CORANK Method's Extractive Summary: [At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [Alexander Vinokourov of Kazakhstan has won stage 13 of the 2007 Tour de France.] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [In the third and longest stage of the Tour de France today, the Swiss leader Fabian Cancellara didnt wait to compete with the sprinters, and used his time trial qualities to just hold them off and haul in a second victory.] [Davide Cassani, a former Tour de France cyclist and now a TV commentator for Italy, said that he had seen Rasmussen in the Dolomites area of the Alps on both June 13 and 14, 2007.] [London, England — For the second time in two days, the sun and the crowds came out to welcome the Tour de France to London.] [The route, which will run through the Greenwich Millennium Village, is being prepared for the Départ.] [Cancellara wins 7.9 km time trial prologue.] []Fabian Cancellara.] [Before todays stage, which he won, Rasmussen said: I want to make absolutely clear Ive had out-of-competition tests prior to the Tour de France, 14 tests during the Tour and all the results are negative ...] [Tour de France:]

CCLTS.MSC Method's Compressive Summary: [London and the south east for the two days , the sun and the crowds came out to welcome the tour de france to london .] [alexandre vinokourov of kazakhstan has won the 2007 tour de france.] [Current yellow jersey, Michael Rasmussen has been withdrawn from the 2007 Tour de France by his team Rabobank for lying to the team about his whereabouts in June, saying he was in Mexico while he was, in fact, in Italy.] [In the third and longest stage of the Tour de France today, the Swiss leader Fabian Cancellara didnt wait to compete with the sprinters, and used his time trial qualities to just hold them off and haul in a second victory.] [Davide Cassani, a former Tour de France cyclist and now a TV commentator for Italy, said that he had seen Rasmussen in the Dolomites area of the Alps on both June 13 and 14, 2007.] [tour de france : yellow jersey rasmussen withdrawn.] [At least half a million onlookers turned out to line the route as the Tour de France Prologue closed the streets of Central London for a day.] [London is preparing for Le Grand Départ of the 2007 Tour de France.] [The parts of the Tour that London will be hosting are the Prologue and Stage One.] [Before todays stage, which he won, Rasmussen said: I want to make absolutely clear Ive had out-of-competition tests prior to the Tour de France, 14 tests during the Tour and all the results are negative …] [One day until Le Grand Départ Friday, July 6, 2007.]

Figure 7: Existing methods, CoRank and CCLTS.MSC extractive and compressive summaries, respectively, obtained after re-executing the approaches.

populations (step-13) takes constant time ($\mathcal{O}(1)$); **(d)** non-dominated sorting (step-14) takes $\mathcal{O}(M(2\mathbb{Z})^2)$ time and the crowding distance operator (step-15) takes $\mathcal{O}(M(2\mathbb{Z})\log(2\mathbb{Z}))$ time (Deb et al., 2002) as there can maximum $2\mathbb{Z}$ solutions (twice of the number of solutions in the population) in a front in a worst-case scenario. Here, $2\mathbb{Z}$ is there because for each solution, there are two new solutions. Let's assume $\mathbb{V} = 2\mathbb{Z}$. Further, the time complexity of sorting the solutions in a front based on the crowding distance operator is $\mathcal{O}((\mathbb{V})\log(\mathbb{V}))$. The overall time complexity of non-dominated sorting and crowding distance calculation is $\mathcal{O}(M\mathbb{V}^2)+\mathcal{O}(M\mathbb{V}\log(\mathbb{V})+\mathcal{O}((\mathbb{V})\log(\mathbb{V}))$ which can be written as $\mathcal{O}(M\mathbb{V}^2)$ after solving it.

Steps-7 to 15 are executed for $\mathbb{G}$ number of generations; therefore, the complexity of theses step will be the sum of (b), (c), and (d), multiplied by $\mathbb{G}$, *i.e.*, $\mathbb{G}(\mathcal{O}(\mathbb{V}M) + \mathcal{O}(1) + \mathcal{O}(M\mathbb{V}^2))$. Thus, the total worst-case complexity of our approach will be

$$\mathcal{O}(\mathbb{Z} + \mathbb{Z}\mathbb{M})+ \mathbb{G}(\mathcal{O}(\mathbb{V}M) + \mathcal{O}(1) + \mathcal{O}(M\mathbb{V}^2))$$
$$\implies \mathcal{O}(ZM) + \mathbb{G} \times \mathcal{O}(M\mathbb{V}^2) \equiv \mathcal{O}(\mathbb{G}M\mathbb{V}^2)$$

As M and $\mathbb{G}$ is constant in our approach; therefore, we can simply say it as $\mathbb{V}^2$.

2.5 Conclusions on Extractive-Compressive Cross-Language Summarisation

We proposed a cross-language summarisation system, namely, CLTS-MEA, where the target language is different from the source language. The problem is treated as a subset of the sentence selection problem. To aid this algorithm, the strength of a multi-objective evolutionary algorithm is utilised where three different perspectives, measuring diversity among sentences, the mutual information of both languages, and relevance of longer sentences, are simultaneously optimised to obtain a good quality of the summary. As it is challenging to decide which perspectives to be used together; therefore, the ablation study is presented by varying the different combinations. Further, for the used datasets, the comparison between the generated extractive and compressive summaries are shown using an example as well as in terms of the ROUGE measure. And, it has been found that using the all mentioned objectives together helps in gaining the improvement of 3.53%, 13.80%, 7.83% in terms of average ROUGE-1, ROUGE-2, and ROUGE-SU4, respectively, over the existing approaches. The readability of summaries was also studied using three different scenarios.

A paper based on this work is currently submitted and under review.

3 Abstractive Cross-lingual Text Summarisation

The abstractive neural summarisation approaches use similar deep learning architectures as machine translation (MT), but face some additional problems: the input is usually longer, the output is short compared to the input, and the content compression is lossy. Current abstractive summarisation may

suffer from repetitive outputs (n-gram repetition), absurd content (creating meaningless sentences and phrases), misrepresented facts (e.g., who won the football match), problems with out-of-vocabulary words (applies to models without a copy mechanism which omit many proper names), or poor content selection (especially for longer texts). Nevertheless, the returned summaries are often useful and of good quality.

Many summarisation approaches exist for resource-rich languages (Aksenov et al., 2020; Scialom, Dray, Lamprier, Piwowarski, & Staiano, 2020). Existing cross-lingual approaches address the problem of a document in one language and its summary in another language, typically English or Chinese (Zhu et al., 2019; Ouyang, Song, & McKeown, 2019), while we are interested in the cross-lingual transfer of trained summarisation models from resource rich-languages to less-resourced languages, i.e. to produce summaries in a less-resourced language. In classification, cross-lingual embeddings present a promising approach for less-resource languages and enable the model transfer from resource-rich to less-resourced languages (Adams, Makarucha, Neubig, Bird, & Cohn, 2017; Artetxe, Ruder, & Yogatama, 2019; Martinc, Pollak, & Robnik-Sikonja, 2021). Typically, this is done by multilingual models such as BERT (Devlin, Chang, Lee, & Toutanova, 2018), or training the model on the resource-rich language (using monolingual embeddings in the source language) and then applying it to the less-resourced language where the input embeddings in the target language are mapped to the source language embeddings. Unfortunately, this standard procedure does not work for cross-lingual summarisation, as the model is trained to output the sentences in the grammar of the source language. Blindly applying the procedure to a summarisation model trained on English would produce sentences with English grammar inthe target language. It is possible to achieve cross-lingual summarisation using translation, but for summarisation, this approach is unsatisfactory, as our baseline models show.

In the proposed solution, we use a pretrained English summarisation model, proposed by Chen and Bansal (2018), and use English as the source language and Slovene as the less-resourced target language. Using cross-lingual embeddings, we map Slovene word embeddings into the English word vector space. As zero-shot transfer learning is not satisfactory, we further fine-tune the resulting model. Our cross-lingual models are trained with increasingly large portions of the available target language dataset. In the output stage of our models, we generate several hypotheses and selected the best one using four evaluation metrics, including a transformer-based neural language model in the target language. Our main contribution is the cross-lingual methodology that produces a useful summarisation model for a less-resourced language. The automatic metrics show that the created summariser is on par with a summarisation model trained from scratch on the target language. In a zero-shot transfer, our cross-lingual approach does not require any resources in the target language apart from a monolingual corpus to build a language model. In a few-shot transfer, a moderate amount of summaries in the target language greatly improve the outputs.

3.1 Architecture of Cross-Lingual Summariser

The proposed approach consists of several steps, presented in Figure 8. Here we describe them step-by-step.

As a pretrained summarisation engine (step 1), we could use several pretrained summarisation models, but in this work, we used the English summariser (Chen & Bansal, 2018). To adapt it to cross-lingual setting, we first replaced the English word embeddings at its input with Slovene embeddings (step 2). To match the word semantics of the two languages, we used the cross-lingual Procrustes alignment (Lample & Conneau, 2019) and mapped the Slovene word embeddings into the English vector space. This already allows us to put Slovene text on the input of the summarisation model (step 3). We fine-tuned the model with different amounts of Slovene text. In step 4, we used the trained model to generate several hypotheses, and in step 5, we assessed the hypotheses to choose the final output. This assessment used an independently trained Slovene language model using transformer architecture and two different metrics. The best hypothesis was included into a summary.

Each of these steps is described in detail in Appendix A.

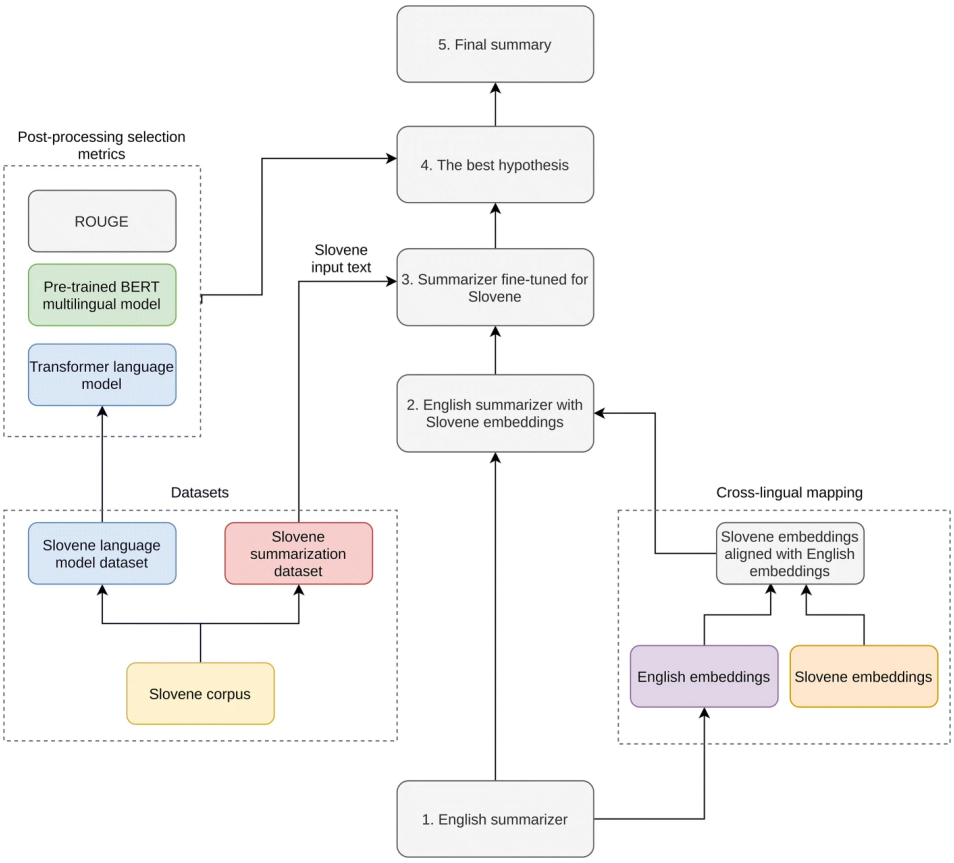


Figure 8: The outline of the proposed cross-lingual summarisation approach.

3.2 Creation of the Final Summary

The English summarisation model is fine-tuned to produce Slovene summaries. Nevertheless, the outputs are sometimes of low quality. For example, sometimes summarisation models produce repeating n-grams, which we eliminate with a rule-based approach. To improve the quality of summaries, we extracted a large number of hypotheses from the abstractive network and assessed them with different heuristics. In the search for hypotheses, we expanded the beam size from standard 4-16 to 64. The heuristic for the assessment of hypotheses consists of two components that try to capture the presence of relevant contents and the readability of hypotheses.

Relevant content. The quality of the content is assessed with two scores. ROUGE score is the standard metric for summarisation quality (Lin & Hovy, 2002) and uses weighted number of matching n-grams between the reference summary and hypothesis. Recently proposed BERTScore (H. Zhang, Xu, & Wang, 2019) is based on the similarity of sentence representation with the pretrained multilingual BERT model (Devlin et al., 2018). We calculated the ROUGE and BERTScore scores by comparing the generated hypotheses from the abstractor network with the sentences extracted with the extractive network.

Text readability. The readability of the generated hypotheses is assessed with two measures: the internal evaluation of hypotheses with the loss function computed by the abstractive neural network, and the Slovene language model. The latter is expressed with the perplexity score, computed as the average entropy per character expressed in bits.

With this approach, we get four different assessments for each generated hypothesis. We first used only one heuristic to select the best hypothesis and analyzed the results. After that, we considered

Table 5: Comparison of our best model with related Slovene model and state-of-the-art English models

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Zidarn (2020) Slovene LSTM	23,77	7,97	23,95	0,679
Our M100 + ROUGE-L and BERTScore	24,97	7,43	21,50	0,679
English Chen and Bansal (2018)	40,88	17,80	38,54	-
English Zhang et al. (2020) - PEGASUS	44,17	21,47	41,11	-

combinations of two heuristics. For example, we first used the ROUGE scores to narrow down the selection to 32 best hypotheses. These 32 hypotheses were scored anew by the language model and the best one according to the language model scores was selected. In combinations of two metrics, we did not require that they belong to different categories, i.e. we allowed a combination of two content-based heuristics or two-readability based heuristics.

3.3 Evaluation Results

In this section we provide an over view of the results, these are provided in detail in Appendix A.

Several models were created, MENG is the baseline zero-shot transfer model, which means that no target language data was used, only the English embeddings were swapped with the mapped Slovene embeddings. The models M1, M10, M25, M50, and M100 use 1%, 10%, 25%, 50%, or 100% of our target language training set (see Section 3) to fine-tune the English model. We also trained the extractor part of each model because only the reinforcement learning optimized extractor was provided by (Chen & Bansal, 2018). Simultaneously, we updated the weights of the pretrained abstractor and, in the final step, optimized the models with the RL component. MSLO is not a cross-lingual model and was trained on the complete target language training set from scratch. Note that the training set in the target language is significantly smaller than the training set in the source language (117,563 summaries for MSLO vs 287,226 for MENG).

The models were compared using the following metrics. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores are the most commonly used metrics in the evaluation of automatically generated text summaries (Lin & Hovy, 2002). It measures the quality of a summary based on the number of overlapping units (n-grams, sequences of texts, etc.) between reference summaries (created by humans) and automatically generated summaries. The most commonly used metrics are ROUGE-N and ROUGE-L. ROUGE-N measures the overlapping of n-grams, e.g., ROUGE-1 for unigrams and ROUGE-2 for bigrams. ROUGE-L measures the longest common subsequence found in the compared summaries. As baseline summarisation models, we use monolingual models and translation-based models. MSLO is a monolingual model, trained on the complete Slovene STA dataset.

Baseline is a purely extractive model that is part of the MSLO model. The third baseline monolingual model is PG, an end-to-end abstractive model (See, Liu, & Manning, 2017). PG is a hybrid between a seq2seq attention model based on LSTMs and pointer networks (Vinyals, Fortunato, & Jaitly, 2015) that enable the model either to copy words via pointing or generate them from a fixed vocabulary. This helps to solve the problem of out-of-vocabulary words. PG uses a coverage mechanism to mitigate the problem of repetition of seq2seq models by preventing the model to focus on the same locations all the time. To establish the MT baseline, we used the Google MT service. We translated the test set from Slovene into English and generated English summaries with the pretrained monolingual English summariser. After that, we translated the generated English summaries back to Slovene.

The comparison showed M100 (cross-lingual model) and MSLO (trained from scratch) are the best models for the Slovene summarisation. These two models use the same amounts of training data. With manual inspection, we were unable to conclude which model is better in terms of readability. However, M100 consistently shows better ROUGE scores: ROUGE-1 is improved for 0.60, ROUGE-2 for 0.19, and ROUGE-L for 0.52. This shows that our cross-lingual approach produces better summaries compared to monolingual models even without additional sentence selection mechanism analysed.

We compare our best summarisation model (M100 + ROUGE-L and BERTScore) to other existing summarisation models for English (as an upper bound of existing technologies) and Slovene. Table 5 shows the results reported by authors of the listed models. In addition to the standard ROUGE scores, we also provide BERTscore where possible. The reported scores are not directly comparable but give a general picture of the success of the proposed cross-lingual approach. The only other neural summarisation model for Slovene was built by (Zidarn, 2020) who used a two-layer LSTM neural network with the attention mechanism, copy mechanism, and beam search. The dataset of this model is the same STA news dataset extracted from Gigafida corpus, but the author uses different train, test, and validation splits. Our model scored higher on ROUGE-1 (1.20 difference) but lower on ROUGE-2 (0.54) and ROUGE-L (2.45). The BERTScore results of both models are identical. Given the existing sources of variation (different subsets of the original data, different splits, and the problematic nature of automatic summary evaluation metrics), we can conclude that both models perform similarly.

In addition we performed a human evaluation of our best model and the best model of Zidarn (2020). For both models, human reported scores of the generated and reference summaries are presented. Both models produce acceptable readability scores, but in terms of accuracy, it seems that our model generates more accurate content. As the bottom part of Table 6 shows, neither cross-lingual nor monolingual Slovene models can compare to English models in terms of performance. English models are usually trained either on the 4 million instances of the Gigaword dataset or the 290k CNN/Daily Mail dataset, which is similar but larger than our Slovene dataset. The English model used in our experiments (Chen & Bansal, 2018) achieves scores that are almost twice as high compared to our Slovene model. Its results are less misleading and mostly represents facts and information accurately. Many manually inspected summaries show that it omits less important dependent clauses. In our model, this behaviour is less frequent.

PEGASUS (J. Zhang, Zhao, Saleh, & Liu, 2020) is currently one of the best abstractive summarisation models. It is based on the transformer neural architecture and presents an interesting novel insight: models are fine-tuned faster and more successfully if they are pretrained on tasks similar to the final task. Authors thus propose two pretraining objectives. One is the BERT masked language model known from (Devlin et al., 2018). Another is the gap sentence generation that selects and masks whole sentences from documents, and concatenates the gap-sentences into a pseudo-summary. The model is pretrained on two very large corpora. The C4 dataset consists of texts from 350M web-pages (750GB). The HugeNews dataset is even larger with 1,5B articles (3,8TB). The model achieved state of the art performance on 12 summarisation tasks.

Additional detail on all of the evaluation steps, and a detailed discussion of the strengths and weaknesses of the summarisation outputs, can be found in Appendix A.

3.4 Conclusions on Cross-lingual Abstractive Summarisation

We developed a neural cross-lingual approach to abstractive summarisation. Our solution is based on using a pretrained model in the resource-rich language (English), whose outputs are fine-tuned to the target language (Slovene) and further refined with sentence selection heuristics. We first showed that zero-shot transfer is unsatisfactory due to its output following the grammar of the source language. In few-shot transfer, we tested how different amounts of training data in target language used in fine-tuning affects the model and discovered that even small amounts of data in the target language significantly improve the quality of produced summaries. Nevertheless, the quantity and quality of the training sets play a huge role, and the target language dataset (Slovene) is not competitive in either respect. This is most evident when analysing diverse topics from the Slovene dataset, where better represented topics are better summarized compared to less represented ones. In addition to the automatic evaluation, we manually analyzed the quality of the results and also conducted a small-scale human evaluation. The assessments show that the accuracy and readability of the generated summaries are acceptable. Two additional contributions of our work are the first Slovene summarisation dataset consisting of news articles, and publicly available character-based transformer neural language model.

4 Temporal Topic Summary Visualisation

Many text corpora, such as news articles, are temporal in nature, with the individual documents distributed across a span of time. As the size and availability of text corpora have continued to increase in recent years, effective analysis of the content of corpora has become challenging. Taking the temporal nature of most corpora into account when analysing the text, makes it more difficult to describe the corpora and to interpret intuitively the results of analysis.

Topic modelling techniques, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), have been used to automatically generate topic groups in text corpora. The generated topics can help in understanding the contents of a corpus by using keywords and topic association probabilities generated by the topic modelling technique. However, interpreting the results of the techniques is not always easy, and the results can seem counter-intuitive when looking only at the weighted keyword lists. Therefore, visualisation techniques have been used extensively to help with the interpretation of the large number of topics generated by these models. The same is true of temporal topic modelling techniques, such as Dynamic Topic Modeling (Blei & Lafferty, 2006), which require additional visualisation techniques to aid intuitive understanding of the temporal segmentation of the topics and their related keywords.

We propose *TeMoTopic* as a contribution to the collection of visualisation techniques for exploring the temporal distribution of topics in text corpora through the use of temporal mosaics. *TeMoTopic* adopts a space-filling approach to show topic distribution over time, and presents keywords related to each topic at the overview level of the visualisation. The visualisation is interactive and, in contrast to many other techniques, enables direct investigation of the source documents associated with individual topics and keywords. This allows the user to get a general sense of the meaning of a topic through its associated keywords, as well as providing the ability to dive into the details of the related documents.

This work is and extension of the *TeMoCo* visualisation which was described in Deliverable D4.3. The system is described in (Sheehan, Luz, & Masoodian, 2021), see Appendix B.

4.1 Topic Visualisation Tasks

The design of a visualisation tool should clearly be motivated by concrete tasks relevant to the end-users of the intended tool. Munzner's *nested model for visualisation design and validation* (Munzner, 2009) describes steps that can be taken to mitigate threats to the validity of a visualisation design. The first of the four levels of this design model is the characterization of domain specific tasks which should be supported by the visual encoding.

Table 6: visualisation tasks for topic model exploration.

Task	Description
Overview of Topics	Visualize topic in terms of extracted keywords
Overview of Document - Topic Relations	View documents related to a topic
Remove Topics from the visualisation	Topic removal from overview
Filtering Documents	View a subset of documents for a topic
Perform Set Operations	Enable exclusion/inclusion of documents in the corpus
Show and Cluster Similar Topics	Enable identification of similar topics
Perform Cluster Operations	Enable grouping of similar topics
Annotating Topics	Allow for labelling of the topics
Visualize Topic Change	View topic distribution and keywords over time

Ganesan, Brantley, Pan, and Chen (2015) identify the key tasks in the design description of *LDAExplore*, which should be supported by visualisations that aim to help users explore the results of Latent Dirichlet Allocation (LDA). Since LDA is one of the most commonly used topic modelling techniques for text corpora, these key tasks could be generalized to other techniques where a corpus is also split into topics, and keywords associated with those topics are extracted.

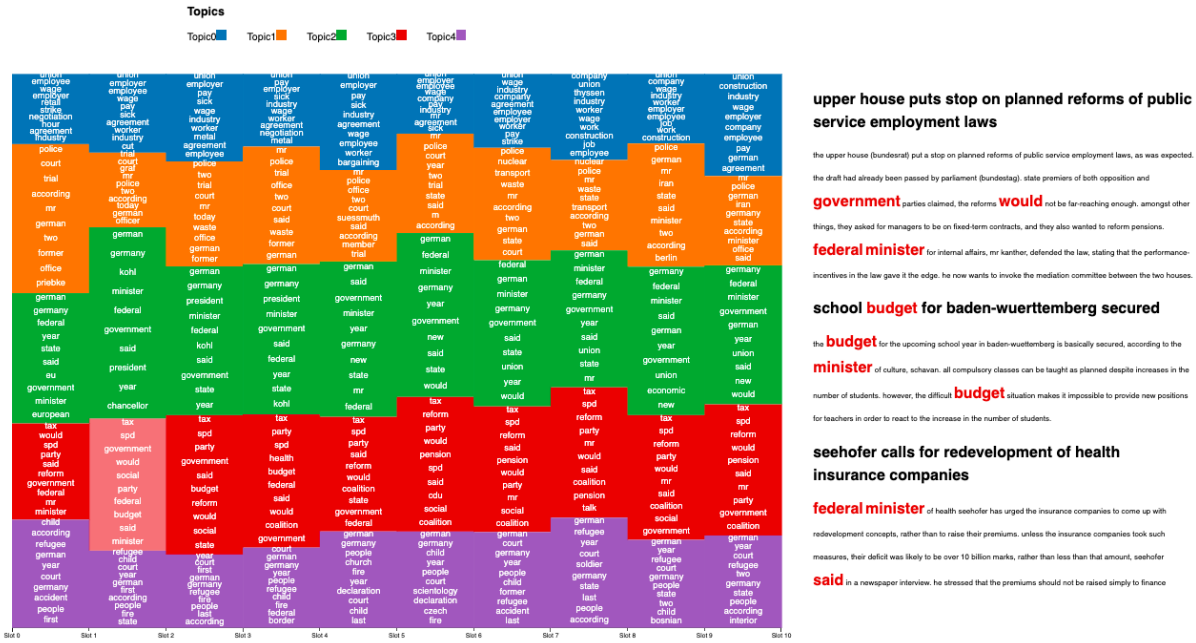


Figure 9: *TeMoTopic* visualisation, showing the temporal mosaic view (left) and the document view (right), showing the selected keywords for the red topic in the second timeslice (red tile on the bottom left).

In addition, Ganesan et al. (2015) argue that the results of LDA can be counter-intuitive, and that the ability to explore and interact with the document set should make the topic and word distributions more intuitive and insightful. Table 6 shows the eight tasks identified by Ganesan et al. (2015), as well as one additional task which we consider to be important for visualizing temporal topics. The table also includes a brief description of the tasks which are fully described by Ganesan et al. (2015).

These tasks describe a need for topic overview with document detail available on-demand, this follows the well-known visual information seeking mantra proposed by Shneiderman (1996). Interactions around viewing, filtering, removing, and combining topics and documents should also be supported. Finally, we include an additional task for visualizing topic changes over time. This modifies the *Overview Topics* task, such that the change in distribution and keywords across is available to explore.

Figure 10: *TeMoTopic* visualisation, showing the temporal mosaic view (left) and the filtered document view (right), with the word ”german“ selected from a temporal topic timeslice (orange tile on the top left).

4.2 TeMoTopic: Temporal Mosaic Topic visualization

Figure 9 shows the *TeMoTopic* visualization tool. It consists of two juxtaposed views (Javed & Elmqvist, 2012): the temporal mosaic (left), and the document view (right). The design of the temporal mosaic is

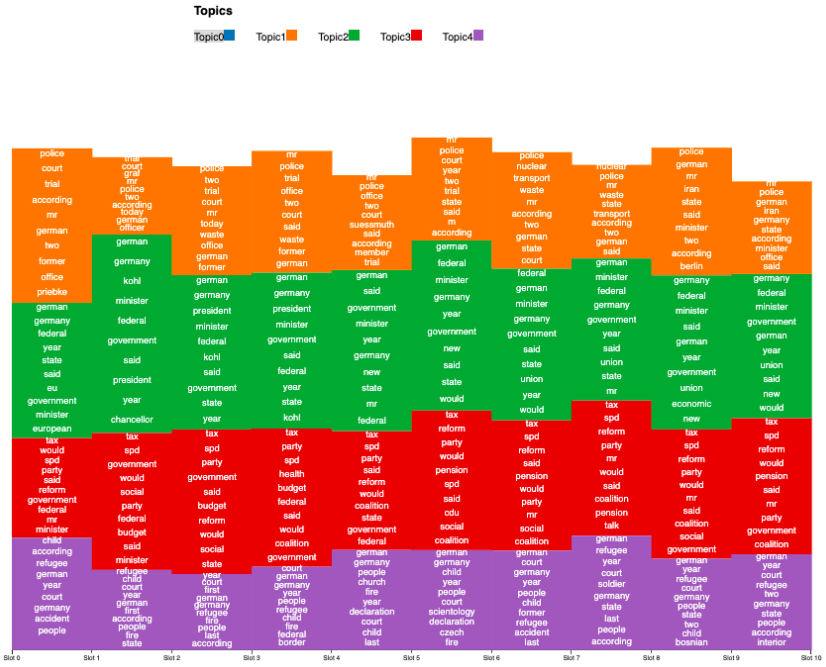


Figure 11: *TeMoTopic* filtered temporal mosaic view after the blue topic was selected for removal via clicking on the legend.

based on a visualization proposed by (Luz & Masoodian, 2007), and further expanded in our previous temporal mosaic visualizations *TeMoCo* visualization (Sheehan, Albert, Masoodian, & Luz, 2019) and *TeMoCo-Doc* visualization (Sheehan, Luz, Albert, & Masoodian, 2020), which have been used to link transcripts of meetings to document reports in a medical context.

TeMoTopic is designed for interactive exploration of temporal topic summaries, each tile in the visualisation represents a temporal topic. The height of the tile represents its topic distribution in the time slice. Keywords associated with the topic are displayed for each timeslice tile. Click interaction on the tiles retrieve the associated documents from that time and topic. The retrieved documents are highlighted and displayed in the accompanying document view 9. Other interactions allow for the retrieval of the documents associated with a single keyword from a top time slice tile (Figure 10) and the removal of entire topics by clicking on the legend (Figure 11).

This visualisation design is motivated by visualisation design theory and the motivation and design decisions are described in greater detail in Appendix B (Sheehan et al., 2021).

4.3 Implementation

The visualization tool[8] is implemented as a single-page web application using the *D3.js* framework (Bostock, Ogievetsky, & Heer, 2011). It takes two JavaScript Object Notation (JSON) files as input: the first file contains topic, keyword, timeslice, weights, and associated filenames, and the second input file is simply a JSON structure containing the documents with filename used as the retrieval key. Sample Python scripts are provided for generating topics and keywords on the sample dataset and for preparing the visualization input files from the model output.

The tool can be used with any temporal topic model but in this example we make use of dynamic topic modelling (Blei & Lafferty, 2006) to identify temporal topics and keywords in a subset of the *de-news*[9]

[8] The software and working example are available at `https://github.com/sfermoy/TeMoCo`.

[9] `http://homepages.inf.ed.ac.uk/pkoehn/publications/de-news/`

corpus of German-English parallel news. The dataset consists of transcribed German radio broadcasts which were manually translated into English. Between 1996 and 2000 volunteers selected and transcribed five to ten of these news broadcasts per day and added them to the dataset. In the examples of *TeMoTopic*, shown in Figures 9, 10, and 11, we selected a ten month span of the dataset and presented the four largest topics. The choice of time span and topic number was only for presentation and to exemplify the interface features. We did not attempt to choose a time period or number of topics based on prior knowledge of the news relevant at the time in Germany. We present our examples to describe the interface and interactions, rather than as an analysis of the dataset, and we choose to draw no conclusions about the dataset contents and topics.

4.4 Conclusions on Temporal Visualisation of Topics

While many other temporal visualisation techniques, such as ThemeRiver (Havre, Hetzler, Whitney, & Nowell, 2002), offer some of the functionality for temporal visualisation of topics or visualisation of content changes, they do not feature implicit linking between the visualisation and the underlying content documents. We consider this to be the main contribution of *TeMoTopic* visualisation and its distinguishing feature with regards to the state of the art.

The paper associated with this work is included in Appendix B (Sheehan et al., 2021)

5 Conclusions

In this report we presented the work performed during the second half of Task 4.2.

We developed cross-language summarisation system, CLTS-MEA, where the target language is different from the source language. The comparison between the generated extractive and compressive summaries are shown using an example as well as in terms of ROUGE measure. And, it has been found that using the all mentioned objectives together improves scores by 3.53%, 13.80%, and 7.83% in terms of average ROUGE-1, ROUGE-2, and ROUGE-SU4, respectively, over the existing approaches.

We developed a neural cross-lingual approach to abstractive summarisation. The solution is based on using a pretrained model in a resource-rich language (English), and fine-tuning the outputs to the target language (Slovene) and further refining them using sentence selection heuristics. Through both metric based and human evaluations we found that the resulting generated summaries are readable and accurate when even small amounts of training data is available. Two additional contributions of this work are the first Slovene summarisation dataset consisting of news articles, and publicly available character-based transformer neural language model.

Our prior work on temporal corpus visualisation was extended to produce the TeMoTopic system, which is publicly available. The system enables exploration of the temporal topic trends and associated keywords in a collection of documents, while enabling the interactive exploration of the visually summarised documents.

6 Associated Outputs

The work described in this deliverable has resulted in the following resources: (Sheehan et al., 2021)

Description	URL	Availability
Code for CLTS-MEA	`https://github.com/jgmorenof/CLTS-MEA`	To become public*
Code for Abstractive Summarisation	`https://github.com/EMBEDDIA/cross-lingual-summarization`	Public (MIT)
Code for TeMoCo	`https://github.com/EMBEDDIA/TeMoCo`	Public (MIT)

* The code will become public after an associated publication is published.

Parts of this work are described in detail in the following publications, which are attached to this deliverable as appendices:

Citation	Status	Appendix
Aleš Žagar and Marko Robnik-Šikonja.Cross-lingual transfer of abstractive summarizer to less-resourced language. Journal of Intelligent Information Systems. 2021	Published	Appendix A
Shane Sheehan, Saturnino Luz and Masood Masoodian.TeMoTopic: Temporal Mosaic Visualisation of Topic Distribution, Keywords, and Context. In the Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. 2021. Association for Computational Linguistics Online	Published	Appendix B

References

Adams, O., Makarucha, A., Neubig, G., Bird, S., & Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (pp. 937–947).

Aksenov, D., Moreno-Schneider, J., Bourgonje, P., Schwarzenberg, R., Hennig, L., & Rehm, G. (2020). Abstractive text summarization based on language model conditioning and locality modeling. *arXiv preprint arXiv:2003.13027*.

Al-Tashi, Q., Abdulkadir, S. J., Rais, H. M., Mirjalili, S., & Alhussian, H. (2020). Approaches to multi-objective feature selection: A systematic literature review. *IEEE Access*, *8*, 125076–125096.

Artetxe, M., Ruder, S., & Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, March). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022.

Bostock, M., Ogievetsky, V., & Heer, J. (2011, December). D3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309.

Chan, L. K., Cheng, S. W., Mead, E., & Panjer, H. (1973). On a t-test for the scale parameter based on sample percentiles. *IEEE Transactions on Reliability*, *22*(2), 82–87.

Chen, Y.-C., & Bansal, M. (2018). Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Deb, K. (2015). Multi-objective evolutionary algorithms. In *Springer handbook of computational intelligence* (pp. 995–1015). Springer.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, *6*(2), 182–197.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, K.-L., & Swamy, M. (2016). Particle swarm optimization. In *Search and optimization by metaheuristics* (pp. 153–173). Springer.

Duan, X., Yin, M., Zhang, M., Chen, B., & Luo, W. (2019). Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3162–3172).

Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd international conference on computational linguistics (coling 2010)* (pp. 322–330).

Ganesan, A., Brantley, K., Pan, S., & Chen, J. (2015). Ldaexplore: Visualizing topic models generated using latent dirichlet allocation. *arXiv preprint arXiv:1507.06593*.

Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., & Varma, V. (2011). Tac 2011 multiling pilot overview.

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). Themeriver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 9-20.

Hussain, T., Muhammad, K., Ullah, A., Cao, Z., Baik, S. W., & de Albuquerque, V. H. C. (2019). Cloud-assisted multiview video summarization using cnn and bidirectional lstm. *IEEE Transactions on Industrial Informatics*, *16*(1), 77–86.

Javed, W., & Elmqvist, N. (2012). Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE* (pp. 1–8). doi: 10.1109/PacificVis.2012.6183556

Jhaveri, N., Gupta, M., & Varma, V. (2019). clstk: The cross-lingual summarization tool-kit. In *Proceedings of the twelfth acm international conference on web search and data mining* (pp. 766–769).

Ladhak, F., Durmus, E., Cardie, C., & McKeown, K. (2020). Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. *arXiv preprint arXiv:2010.03093*.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Li, L., & Li, T. (2013). An empirical study of ontology-based multi-document summarization in disaster management. *IEEE transactions on systems, man, and cybernetics: systems*, *44*(2), 162–171.

Li, W., & Zhuge, H. (2019). Abstractive multi-document summarization based on semantic link network. *IEEE Transactions on Knowledge and Data Engineering*.

Li, X., Du, L., & Shen, Y.-D. (2012). Update summarization via graph-based sentence ranking. *IEEE transactions on Knowledge and Data Engineering*, *25*(5), 1162–1174.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Lin, C.-Y., & Hovy, E. (2002). Automated multi-document summarization in neats. In *Proceedings of the human language technology conference (hlt2002)* (pp. 23–27).

Linhares Pontes, E., Huet, S., Gouveia da Silva, T., Linhares, A. C., & Torres-Moreno, J.-M. (2018, June). Multi-sentence compression with word vertex-labeled graphs and integer linear programming. In *Proceedings of the twelfth workshop on graph-based methods for natural language processing (TextGraphs-12)* (pp. 18–27). New Orleans, Louisiana, USA: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W18-1704` doi: 10.18653/v1/W18-1704

Linhares Pontes, E., Huet, S., Torres-Moreno, J.-M., & Linhares, A. C. (2020). Compressive approaches for cross-language multi-document summarization. *Data & Knowledge Engineering*, *125*, 101763.

Litvak, M., & Last, M. (2013). Cross-lingual training of summarization systems using annotated corpora in a foreign language. *Information retrieval*, *16*(5), 629–656.

Luz, S., & Masoodian, M. (2007, July). Visualisation of Parallel Data Streams with Temporal Mosaics. In *Proceedings of the 11th International Conference Information Visualization* (pp. 197–202). doi: 10.1109/IV.2007.127

Martinc, M., Pollak, S., & Robnik-Sikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, *47*(1), 141–179.

McDonald, D. D. (2010). Natural language generation. *Handbook of Natural Language Processing*, *2*, 121–144.

Mezura-Montes, E., Velázquez-Reyes, J., & Coello Coello, C. A. (2006). A comparative study of differential evolution variants for global optimization. In *Proceedings of the 8th annual conference on genetic and evolutionary computation* (pp. 485–492).

Mills, M. T., & Bourbakis, N. G. (2013). Graph-based methods for natural language processing and understanding—a survey and analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *44*(1), 59–71.

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of student's t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia*, *22*(4), 407.

Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 921-928.

Oliveira, L. S., Sabourin, R., Bortolozzi, F., & Suen, C. Y. (2002). Feature selection using multi-objective genetic algorithms for handwritten digit recognition. In *Object recognition supported by user interaction for service robots* (Vol. 1, pp. 568–571).

Ouyang, J., Song, B., & McKeown, K. (2019). A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 2025–2031).

Pontes, E. L., Huet, S., Torres-Moreno, J., da Silva, T. G., & Linhares, A. C. (2020). A multilingual study of multi-sentence compression using word vertex-labeled graphs and integer linear programming. *Computación y Sistemas*, *24*(2). Retrieved from `https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3335`

Pontes, E. L., Huet, S., Torres-Moreno, J.-M., & Linhares, A. C. (2018). Cross-language text summarization using sentence and multi-sentence compression. In *International conference on applications of natural language to information systems* (pp. 467–479).

Rudra, K., Goyal, P., Ganguly, N., Imran, M., & Mitra, P. (2019). Summarizing situational tweets in crisis scenarios: An extractive-abstractive approach. *IEEE Transactions on Computational Social Systems*, *6*(5), 981–993.

Saini, N., Saha, S., Bhattacharyya, P., & Tuteja, H. (2020). Textual entailment–based figure summarization for biomedical articles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(1s), 1–24.

Saini, N., Saha, S., Chakraborty, D., & Bhattacharyya, P. (2019). Extractive single document summarization using binary differential evolution: Optimization of different sentence quality measures. *PloS one*, *14*(11), e0223477.

Schrijver, A. (1998). *Theory of linear and integer programming*. John Wiley & Sons.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., & Staiano, J. (2020). Discriminative adversarial search for abstractive summarization. In *International conference on machine learning* (pp. 8555–8564).

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Shareghi, E., & Hassanabadi, L. S. (2008). Text summarization with harmony search algorithm-based sentence extraction. In *Proceedings of the 5th international conference on soft computing as transdisciplinary science and technology* (pp. 226–231).

Sheehan, S., Albert, P., Masoodian, M., & Luz, S. (2019, Jun). TeMoCo: A Visualization tool for Temporal Analysis of Multi-Party Dialogues in Clinical Settings. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 690–695).

Sheehan, S., Luz, S., Albert, P., & Masoodian, M. (2020). Temoco-doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents. In *Proceedings of the international conference on advanced visual interfaces.* New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3399715.3399956` doi: 10.1145/3399715.3399956

Sheehan, S., Luz, S., & Masoodian, M. (2021, April). TeMoTopic: Temporal mosaic visualisation of topic distribution, keywords, and context. In *Proceedings of the eacl hackashop on news media content analysis and automated report generation* (pp. 56–61). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.hackashop-1.8`

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages* (pp. 336–343). doi: 10.1109/VL.1996.545307

Vanetik, N., Litvak, M., Churkin, E., & Last, M. (2020). An unsupervised constrained optimization approach to compressive summarization. *Information Sciences*, *509*, 22–35.

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. *arXiv preprint arXiv:1506.03134*.

Wan, X. (2011). Using bilingual information for cross-language document summarization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1546–1555).

Wan, X., Luo, F., Sun, X., Huang, S., & Yao, J.-g. (2019). Cross-language document summarization via extraction and ranking of multiple summaries. *Knowledge and Information Systems*, *58*(2), 481–499.

Wang, B.-C., Li, H.-X., Li, J.-P., & Wang, Y. (2018). Composite differential evolution for constrained evolutionary optimization. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(7), 1482–1495.

Wu, G., Shen, X., Li, H., Chen, H., Lin, A., & Suganthan, P. N. (2018). Ensemble of differential evolution variants. *Information Sciences*, *423*, 172–186.

Yao, J.-g., Wan, X., & Xiao, J. (2015). Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 118–127).

Yao, K., Zhang, L., Du, D., Luo, T., Tao, L., & Wu, Y. (2018). Dual encoding for abstractive text summarization. *IEEE transactions on cybernetics*, *50*(3), 985–996.

Zhang, H., Xu, J., & Wang, J. (2019). Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning* (pp. 11328–11339).

Zhang, J., Zhou, Y., & Zong, C. (2016). Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(10), 1842–1853.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765.

Zhang, Y., Er, M. J., Zhao, R., & Pratama, M. (2016). Multiview convolutional neural networks for multidocument extractive summarization. *IEEE transactions on cybernetics*, *47*(10), 3230–3242.

Zhang, Y., Gong, D.-w., Gao, X.-z., Tian, T., & Sun, X.-y. (2020). Binary differential evolution with self-learning for multi-objective feature selection. *Information Sciences*, *507*, 67–85.

Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., & Zong, C. (2019). Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

Zhu, J., Zhou, Y., Zhang, J., & Zong, C. (2020). Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1309–1321).

Zidarn. (2020). Automatic text summarization of slovene texts using deep neural networks. *In University of Ljubljana faculty of computer and information science, Ljubljana, (MSc thesis, in Slovene)*..

Appendix A: Cross-lingual transfer of abstractive summarizer to less-resource language

Noname manuscript No.
(will be inserted by the editor)

Cross-lingual Transfer of Abstractive Summarizer to Less-resource Language

Aleš Žagar · Marko Robnik-Šikonja

Received: date / Accepted: date

Abstract Automatic text summarization extracts important information from texts and presents the information in the form of a summary. Abstractive summarization approaches progressed significantly by switching to deep neural networks, but results are not yet satisfactory, especially for languages where large training sets do not exist. In several natural language processing tasks, a cross-lingual model transfer is successfully applied in less-resource languages. For summarization, the cross-lingual model transfer was not attempted due to a non-reusable decoder side of neural models that cannot correct target language generation. In our work, we use a pre-trained English summarization model based on deep neural networks and sequence-to-sequence architecture to summarize Slovene news articles. We address the problem of inadequate decoder by using an additional language model for the evaluation of the generated text in target language. We test several cross-lingual summarization models with different amounts of target data for fine-tuning. We assess the models with automatic evaluation measures and conduct a small-scale human evaluation. Automatic evaluation shows that the summaries of our best cross-lingual model are useful and of quality similar to the model trained only in the target language. Human evaluation shows that our best model generates summaries with high accuracy and acceptable readability. However, similar to other abstractive models, our models are not perfect and may occasionally produce misleading or absurd content.

Keywords automatic summarization · text generation · deep neural networks · language models · cross-lingual embeddings · abstractive summarization

A. Žagar and M. Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, Ljubljana, Slovenia.
E-mail: ales.zagar@fri.uni-lj.si, E-mail: marko.robnik@fri.uni-lj.si

1 Introduction

Summarization is a process of extracting or collecting important information from texts and presenting that information in the form of a summary. According to the output of the process, summarization can be broadly divided into an extractive and abstractive type. The extractive approach is non-productive in a sense that it copies important sentences, and the resulting summary does not include new words or sentences. The abstractive approach is creative and produces summaries that rephrase the given content and can contain originally unused words.

The abstractive neural summarization approaches use similar deep learning architectures as machine translation (MT), but face some additional problems: the input is usually longer, the output is short compared to the input, and the content compression is lossy. Current abstractive summarization may suffer from repetitive outputs (n-gram repetition), absurd content (creating meaningless sentences and phrases), misrepresented facts (e.g., who won the football match), problems with out-of-vocabulary words (applies to models without a copy mechanism which omit many proper names), or poor content selection (especially for longer texts). Nevertheless, the returned summaries are often useful and of good quality.

Many summarization approaches exist for resource-rich languages [2, 6, 40]. Existing cross-lingual approaches address the problem of a document in one language and its summary in another language, typically English or Chinese [50, 33], while we are interested in the cross-lingual transfer of trained summarization models from resource rich-languages to less-resourced languages, i.e. to produce summaries in a less-resourced language. In classification, cross-lingual embeddings present a promising approach for less-resource languages and enable the model transfer from resource-rich to less-resourced languages [1, 3, 26]. Typically, this is done by multilingual models such as BERT [13], or training the model on the resource-rich language (using monolingual embeddings in the source language) and then applying it to the less-resourced language where the input embeddings in the target language are mapped to the source language embeddings. Unfortunately, this standard procedure does not work for cross-lingual summarization, as the model is trained to output the sentences in the grammar of the source language. Blindly applying the procedure to a summarization model trained on English would produce sentences with English grammar in the target language. It is possible to achieve cross-lingual summarization using translation, but for summarization, this approach is unsatisfactory, as our baseline models, described in Section 5.1 show.

In the proposed solution, we use a pretrained English summarization model, proposed by Chen and Bansal [9], and use English as the source language and Slovene as the less-resourced target language. Using cross-lingual embeddings, we map Slovene word embeddings into the English word vector space. As zero-shot transfer learning is not satisfactory, we further fine-tune the resulting model. Our cross-lingual models are trained with increasingly large portions of the available target language dataset. In the output stage of our models, we generate several hypotheses and selected the best one using four evaluation metrics, including a transformer-based neural language model in the target language.

Our main contribution is the cross-lingual methodology that produces a useful summarization model for a less-resourced language. The automatic metrics show that the created summarizer is on par with a summarization model trained from scratch on the target language. In a zero-shot transfer, our cross-lingual approach does not require any resources in the target language apart from a monolingual corpus to build a language model. In a few-shot transfer, a moderate amount of summaries in the target language greatly improve the outputs.

The paper is split into further five sections. In Section 2, we present related works, and in Section 3, we describe the Slovene datasets to build the output selection language models and to fine-tune the summarization model in the few-shot transfer experiments. Section 4 outlines the proposed cross-lingual summarization model and gives details of the used components. We report the results in Section 5, and present conclusions and ideas for further work in Section 6.

2 Background and related work

We split this section into three parts. In Section 2.1, we first describe monolingual approaches to text summarization in English and other languages, followed by cross-lingual summarization attempts in Section 2.2. As our approach is based on cross-lingual embeddings, we shortly outline relevant background in Section 2.3.

2.1 Monolingual approaches to text summarization

Most early summarization approaches used the extractive approach also suitable for a multi-document summarization [16]. Lately, deep neural networks learning sequence to sequence (seq2seq) transformations produced state of the art abstractive summaries [39, 31]. Seq2seq models first encode a source document into an internal numeric representation and then decode it into an abstractive summary. These models work best for short single-document summaries, e.g., headline generation and news summarization. They use the attention mechanism which ensures that the decoder focuses on the appropriate input words [5]. They frequently use the copy mechanism that copies relevant words from the input when they are not present in a dictionary [41], and the coverage mechanism that avoids redundant contents [44]. Auxiliary tasks, e.g., keyphrase extraction, can improve the summarization results [27]. Currently, all of the best summarization models [37], [48], [14] are based on the transformer architecure [45].

As we use Slovene as the target language, we report the work on summarization in this language. Recently, Zidarn [51] built the first abstractive summarizer for the Slovene language using the seq2seq architecture and deep neural networks. The best results were produced by a two-layer LSTM with attention mechanism, copy mechanism, and beam search. To allow comparison, we use the same target language dataset of approximately 120,000 news. Zidarn [51] showed that this dataset is not large enough to achieve results comparable to English.

Besides English, there are only a few other languages with abstractive summarizers. Straka et al. [42] presented SumeCzech, a large news summarization dataset for

Czech (1 million samples). For summarization, they compared unsupervised methods such as TextRank [28], returning a few first sentences, and supervised methods (logistic regression and random forests) on handcrafted features. Fecht et al. [15] used the encoder-decoder architecture on German Wikipedia articles (100,000 samples), where the summary is the first section of the article and the subsequent text represents the document. Hu et al. [20] created a Chinese summarization dataset (2.4 million samples) from a Chinese microblogging website Sina Weibo and used a recurrent neural network for abstractive summarization.

2.2 Cross-lingual approaches to text summarization

Most existing cross-lingual summarization attempts aim to obtain a summary in a different language than the original document. For that purpose, they use summarization in combination with MT. Zhu et al. [50] proposed a cross-lingual approach suitable for resource-rich languages where both source and target language have enough training data to build a summarizer. Two different translation schemes are used: ”translate then summarize” scheme first translates the original document into the target language and then generates a summary; ”summarize then translate” scheme first generates a summary and then translates it into the target language. Zhu et al. [50] used the MT on a large English and Chineese corpus to first create a cross-lingual summarization dataset and then trained a neural network in an end-to-end manner incorporating both MT and summarization.

Ouyang et al. [33] aimed at summarizing documents in low-resource languages in the resource-rich language (English). To address the problem of noisy MT from low-resource languages, they translated documents from an English document-summary corpus to three low-resource languages and back into English. They coupled noisy documents with the original summaries and trained the neural network summarization architecture proposed by See et al. [41] on the obtained corpus. The approach was shown to improve over the ”translate then summarize” scheme as the neural network took into account some of the errors introduced by MT from less-resourced languages. In our work, we address a situation where we want to obtain the summary in the same less-resource language as the original text. Our cross-lingual approach uses the pretrained summarization model in the resource-rich language and fine-tunes it to the target language. We use MT as a baseline (translate-summarize-translate), and show that it is not competitive with our direct cross-lingual model transfer approach.

Chi et al. [10] outperformed machine-translation-based approaches in a headline generation task by pre-training a seq2seq transformer model [45] under both monolingual and cross-lingual settings. For the pretraining procedure, they used various tasks: monolingual masked language modeling, denoising auto-encoding objective (to pre-train the encoder-decoder attention mechanism), cross-lingual masked language modeling, and cross-lingual auto-encoding. After the pretraining phase, the model was fine-tuned on question generation and abstractive summarization tasks. In contrast to the headline generation task, where the outputs are short and require little grammar, we work with much longer summaries. To accommodate to less-resourced languages, our approach uses cross-lingual word embeddings at the input to the al-

ready pretrained summarization model and adapts the decoder phase to fit the target language better.

Our cross-lingual approach is based on the monolingual model proposed by Chen and Bansal [9]. This hybrid summarization model first selects salient sentences and then paraphrases them. The model is comprised of two independently trained neural networks bridged by policy-based reinforcement learning. We describe this model in Section 4.2.

2.3 Word embeddings

The idea of word embeddings is to learn high-dimensional vectors that capture the meaning of words. Popular variants are Word2vec [29], GloVe [35], fastText [18], ELMo [36], and BERT [13]. An important insight for our work is that relations between words in the embedded space are preserved across languages [30]. Cross-lingual embeddings align monolingual embeddings into a joint vector space [38]. In the beginning, these techniques required parallel corpora or a bilingual dictionary to map words from a source to target language. Recent approaches can train cross-lingual embeddings in an unsupervised manner [23, 3]. A major drawback of classical word embeddings is that they cannot deal with polysemy. Recent contextual embeddings, ELMo [36] and BERT [13], learn a different representation for each word based on its context.

3 Datasets

We describe the creation of two datasets, one for the summarization task and the other for the language modeling used in the output selection. Both datasets were extracted from the Gigafida corpus [21] of written standard Slovene. The corpus consists of newspapers, magazines, and web texts, and contains 38,310 documents with more than 1.1 billion words. We end the section with a short discussion on approximations to true human summaries used in existing summarization datasets.

3.1 Slovene summarization dataset

The summarization dataset contains news and their summaries from the Slovenian press agency (STA) news web texts. The first paragraph of each news article is taken as a summary and the rest of it as the text of the news. Since the Gigafida corpus from which we extracted STA news is sentence segmented but not paragraph segmented, we designed a heuristic to extract the first paragraph. We started with 284,000 training samples but kept only texts between 1,000 and 3,000 characters. Some texts were discarded as they contained weather reports, lists of events around the world, etc., and some of them were too long. A total of 127,563 samples remained and were split into the train, test and validation set. Both the test and validation set contain 5,000 instances, and the training set contains the remaining 117,563 news.

3.2 Slovene language model dataset

To create our cross-lingual summarization model, we started with the trained English model. While a cross-lingual mapping can transfer the target language (Slovene) into the required input space of the source language (English), this is not sufficient to produce sensible texts in the target language because the grammar of the decoder remains in the source language. Our output modifications require that we train a language model in the target language. For that purpose, we trained a character-level Slovene language model. Bojanowski et al. [7] discovered that language models for morphologically rich languages (such as Slovene) are improved by using character-level information. As the training set we used the Gigafida corpus which is tokenized and sentence segmented. All punctuation, special characters, and numbers were preserved, but alphabetical characters were lower-cased. A total of 59,861,870 sentences were extracted with the average sentence length of 242 characters. The sentences were split into the train, test, and validation set with ratios of 90:5:5.

3.3 Approximations to true human summaries

The aim of our work is to produce methodology for cross-lingual transfer of trained summarizers. To evaluate such a system in zero-shot and few-shot transfer mode, we need a reasonably sized dataset in the target less-resourced language. Unfortunately, there is no such summarization dataset with actual human abstracts in Slovene and our investigation showed that the same is true for other languages, as all existing large datasets use approximations.

The most commonly used English summarization dataset CNN/DM [31] does not contain actual human abstracts but only the main bullet points (highlights). Another widely used English dataset, the Gigaword summarization dataset [17], is a headline generation task. The Newsroom summaries were produced from the metadata available in the HTML pages of articles using various keywords with no standard metadata format [19]. Non-English datasets are produced in a similar way. For example, in the Slovak SME dataset, Suppa and Adamec [43] joined the headline of an article with its lead paragraph to form the target summary.

We could not find any real abstracts of the described datasets for human evaluation purposes. The only datasets we are aware of and contain proper abstracts are too small and not appropriate for neural summarization [24, 34]. If we built a new small Slovene dataset with the actual abstracts, it would not be sufficiently large for training and would not match the properties of the training datasets. Such a small new dataset would introduce a task transfer problem and we would lose the possibility to compare our results with other approaches (e.g., Slovene evaluation in Table 7), which used existing approximating datasets. We believe that this task transfer should be approached in further work and studied carefully.

Large English datasets which do contain actual abstracts are based on much longer texts, e.g., ArXiv and PubMed abstracts [11] or book summaries [22]. These datasets are typically not treated with neural abstractive summarization approaches

used in our work but use an extractive approach or a hybrid extractive-abstractive approach. These approaches are outside the scope of this work.

4 Architecture and implementation of cross-lingual summarizer

In this section, we first outline our solution to the problem of cross-lingual summarization. After that, we provide descriptions of components used: cross-lingual word embeddings for the input, fine-tuning of pretrained English summarization model to Slovene, generation and evaluation of the best hypothesis with several evaluation scores, including the Slovene language model.

4.1 Architecture of the cross-lingual summarizer

The proposed approach consists of several steps, presented in Figure 1. Below we describe them step-by-step.

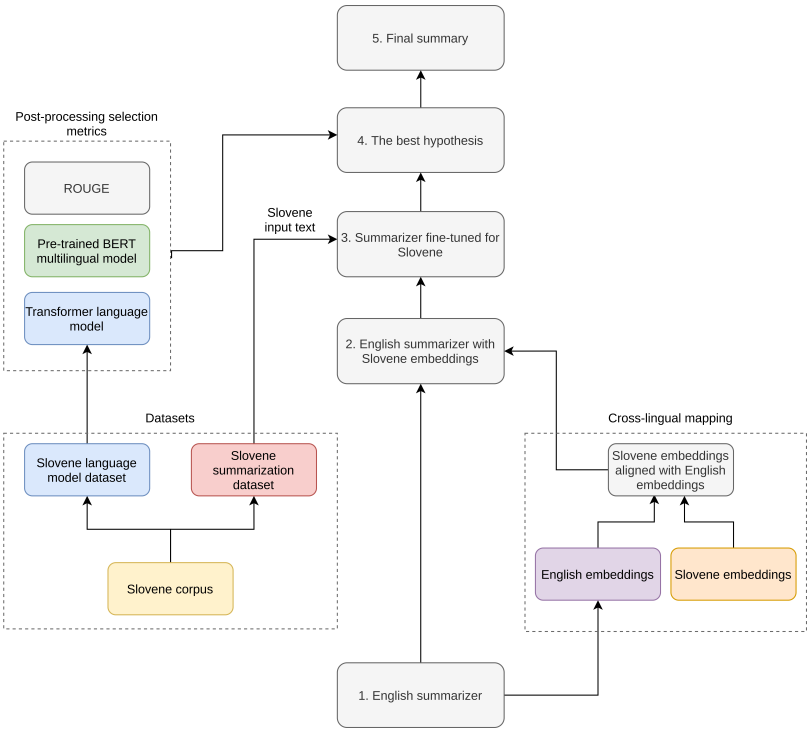


Fig. 1 The outline of the proposed cross-lingual summarization approach.

As a pretrained summarization engine (step 1), we could use several pretrained summarization models, but in this work, we used the English summarizer [9], as described in Section 4.2. To adapt it to cross-lingual setting, we first replaced the English word embeddings at its input with Slovene embeddings (step 2), as described in Section 4.3. To match the word semantics of the two languages, we used the cross-lingual Procrustes alignment [23] and mapped the Slovene word embeddings into the English vector space. This already allows us to put Slovene text on the input of the summarization model (step 3). We fine-tuned the model with different amounts of Slovene text as discussed in Section 4.4. In step 4, we used the trained model to generate several hypotheses, and in step 5, we assessed the hypotheses to choose the final output. This assessment used an independently trained Slovene language model using transformer architecture (described in Section 4.5) and two different metrics, described in Section 4.6. The best hypothesis was included into a summary.

4.2 English summarization model

As our source language summarization model, we used the pretrained summarization model proposed by Chen and Bansal [9]. The model uses customarily trained word2vec embeddings and thus allows a cross-lingual mapping. The architecture of the model is relatively complex and belongs to hybrid approaches to text summarization that combine abstractive and extractive elements. On a high level, it consists of i) the extractive network (that selects salient sentences), ii) the abstractive network (that rewrites or paraphrases them), and iii) the reinforcement learning (RL) step that optimizes the model end-to-end. Both the extractor and abstractor networks are trained independently. During the RL step, the model updates only the extractor weights and leaves the abstractor as it is. The model was trained on the CNN/DailyMail dataset[1], which contains 287,226 training summary/text pairs, 13,368 validation pairs and 11,490 test pairs. The details are available in [9].

4.3 Cross-lingual input alignment

At the input to the neural network summarization model, words are encoded into numeric vectors using word embeddings. In our cross-lingual setting, we use the Slovene input and map it into the English vector space. As the Slovene embedding model, we used the pretrained Slovene fastText embeddings [18], trained on a mixture of Slovene Wikipedia and Common Crawl data[2]. FastText embeddings are constructed with the word2vec CBOW algorithm [29], extended with position weights and subword information. FastText embeddings are especially suitable for morphologically rich languages such as Slovene. To transform Slovene embeddings into the English vector space, we used the MUSE library [23] in a supervised setting. For this transformation, MUSE internally created a train dictionary of size 5,000 and a test dictionary of size 1,500. We replaced the English dictionary with the Slovene

[1] https://cs.nyu.edu/ kcho/DMQA/

[2] https://fasttext.cc/docs/en/crawl-vectors.html

dictionary which was built from 30,000 most common words in the Slovene training dataset. The role of the dictionary is to map words to their embeddings.

4.4 Fine-tuning of the summarization model

Once the target language input (Slovene) is mapped to the source language (English), it is used as an input to the summarization model. Such a model can be already used for summarization in the target language (zero-shot transfer). The resulting summaries adhere to the source language grammar and are of low quality. If any target language summaries are available, we can improve the summarization model with additional training instances (few-shot transfer). To analyze the required amount of additional data, we created several models, presented in Table 1. The models differ in the quantity of additional target language data used in their fine-tuning.

	Slovene dataset size		
Model	in %	# instances	Details
MENG	0%	0	cross-lingual mappings, no fine-tuning, zero-shot transfer
M1	1%	1,176	cross-lingual mappings, trained extractor, fine-tuned abstractor
M10	10%	11,756	cross-lingual mappings, trained extractor, fine-tuned abstractor
M25	25%	29,391	cross-lingual mappings, trained extractor, fine-tuned abstractor
M50	50%	58,782	cross-lingual mappings, trained extractor, fine-tuned abstractor
M100	100%	117,563	cross-lingual mappings, trained extractor, fine-tuned abstractor
MSLO	100%	117,563	Slovene embeddings, trained extractor, trained abstractor, no transfer

Table 1 The produced models using different amounts of target language data (Slovene) in the fine-tuning of the original summarization model.

MENG is the baseline zero-shot transfer model, which means that no target language data was used, only the English embeddings were swapped with the mapped Slovene embeddings. The models M1, M10, M25, M50, and M100 use 1%, 10%, 25%, 50%, or 100% of our target language training set (see Section 3) to fine-tune the English model. We also trained the extractor part of each model because only the reinforcement learning optimized extractor was provided by Chen and Bansal [9]. Simultaneously, we updated the weights of the pretrained abstractor and, in the final step, optimized the models with the RL component.

MSLO is not a cross-lingual model and was trained on the complete target language training set from scratch. Note that the training set in the target language is significantly smaller than the training set in the source language (117,563 summaries for MSLO vs 287,226 for MENG).

4.5 Training the Slovene language model

The adapted and fine-tuned models produce summaries in Slovene, but the quality is not always adequate. For that reason, we used the decoder to generate several hypotheses, post-processed them, and selected the best one according to different evaluation approaches (described below in Section 4.6). As we aim to optimize the

fluency and grammatical correctness of the output sentences, one of the evaluation approaches uses language models. For that purpose, we trained a character-level language model in the target language (Slovene).

Many of the current state-of-the-art language models [4, 12], trained on datasets similar to ours [8], use variants of the transformer architecture [45]. We used the transformer decoder as implemented in the Tensor2tensor library [46], Adam optimizer, 8 attention heads, 6 hidden layers, and position-wise feed-forward networks with one hidden layer of size 2048 and ReLU activation function. These are standard hyperparameters for training on a single GPU. We increased the maximum size of the input from 256 to 512, which is approximately the 95th percentile of the sentence character length in the training corpus. Shorter sentences are padded with spaces and longer are cut off. The dictionary contains 581 characters. The total number of learning parameters was 19,035,136.

The language model was trained for 100 epochs in two parts due to limited computational resources: 60 epochs using 30 million sentences, and 40 epoch with another 23 million sentences). The batch size was 2048. The model was evaluated on the test set with 10k sentences (see Section 3). Training took approximately 4 days on Nvidia Titan X 12GB GPU.

4.6 Creation of the final summary

The English summarization model is fine-tuned to produce Slovene summaries. Nevertheless, the outputs are sometimes of low quality. For example, sometimes summarization models produce repeating n-grams, which we eliminate with a rule-based approach. To improve the quality of summaries, we extracted a large number of hypotheses from the abstractive network and assessed them with different heuristics. In the search for hypotheses, we expanded the beam size from standard 4-16 to 64. The heuristic for the assessment of hypotheses consists of two components that try to capture the presence of relevant contents and the readability of hypotheses.

Relevant content. The quality of the content is assessed with two scores. ROUGE score is the standard metric for summarization quality [25] and uses weighted number of matching n-grams between the refrence summary and hypothesis. Recently proposed BERTScore [49] is based on the similarity of sentence representation with the pretrained multilingual BERT model [13]. We calculated the ROUGE and BERTScore scores by comparing the generated hypotheses from the abstractor network with the sentences extracted with the extractive network.

Text readability. The readability of the generated hypotheses is assessed with two measures: the internal evaluation of hypotheses with the loss function computed by the abstractive neural network, and the Slovene language model described in Section 4.5. The latter is expressed with the perplexity score, computed as the average entropy per character expressed in bits.

With this approach, we get four different assessments for each generated hypothesis. We first used only one heuristic to select the best hypothesis and analyzed the results. After that, we considered combinations of two heuristics. For example, we

first used the ROUGE scores to narrow down the selection to 32 best hypotheses. These 32 hypotheses were scored anew by the language model and the best one according to the language model scores was selected. In combinations of two metrics, we did not require that they belong to different categories, i.e. we allowed a combination of two content-based heuristics or two-readability-based heuristics.

5 Evaluation

In this section, we provide the results and analyses of the created summarization models. We start with the presentation of the evaluation metrics and baseline models in Section 5.1. The results of baseline and fine-tuned models are presented in Section 5.2. The best of the fine-tuned models is further analyzed in Section 5.3 where we compare the proposed heuristics for the selection of output sentences. Section 5.4 contains the human evaluation of the best-produced model. We compare our results with the related approaches in Section 5.5. Finally, in Section 5.6, we manually analyze strengths and weaknesses of our best model.

5.1 Evaluation metrics and baselines

We first present the standard evaluation metrics used in summarization, ROUGE-1, ROUGE-2, and ROUGE-L. Next, we describe the baseline models, both monolingual and translation-based.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores are the most commonly used metrics in the evaluation of automatically generated text summaries [25]. It measures the quality of a summary based on the number of overlapping units (n-grams, sequences of texts, etc.) between reference summaries (created by humans) and automatically generated summaries. The most commonly used metrics are ROUGE-N and ROUGE-L. ROUGE-N measures the overlapping of n-grams, e.g., ROUGE-1 for unigrams and ROUGE-2 for bigrams. ROUGE-L measures the longest common subsequence found in the compared summaries.

As baseline summarization models, we use monolingual models and translation-based models. MSLO is a monolingual model, trained on the complete Slovene STA dataset. EXT Baseline is a purely extractive model that is part of the MSLO model. The third baseline monolingual model is PG, an end-to-end abstractive model [41]. PG is a hybrid between a seq2seq attention model based on LSTMs and pointer networks [47] that enable the model either to copy words via pointing or generate them from a fixed vocabulary. This helps to solve the problem of out-of-vocabulary words. PG uses a coverage mechanism to mitigate the problem of repetition of seq2seq models by preventing the model to focus on the same locations all the time.

To establish the MT baseline, we used the Google MT service. We translated the test set from Slovene into English and generated English summaries with the pretrained monolingual English summarizer. After that, we translated the generated English summaries back into Slovene.

5.2 Results of cross-lingual fine-tuning

As described in Section 4.4, the pretrained summarization model can be improved with different amounts of training data in the target language. Table 2 shows the results of the six models listed in Table 1 and the three baseline monolingual models, described in Section 5.1.

	Average generated		Evaluation scores			
Model	sentences	characters	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity
MENG	3,99	500,61	18,91	3,74	16,27	3,69
M1	2,81	218,48	12,94	1,96	11,61	4,23
M10	1,95	204,59	15,71	3,71	13,87	2,14
M25	2,89	159,00	19,32	5,00	17,12	2,19
M50	3,01	168.59	21,30	6,09	18,91	2,15
M100	2,79	297,67	21,67	6,81	19,16	**2,12**
MSLO	2,58	270,79	21,07	6,62	18,64	2,13
MT Baseline	4,02	297,06	19,76	3,64	17,14	4,26
EXT Baseline	2,58	510,37	22,71	5,58	18,46	/
PG [41]	1,79	270,73	**23,57**	**7,76**	**20,04**	3,15
Reference Slovene	2,10	302,02				
Reference English	3,88	312,51				

Table 2 The performance of the cross-lingual models with different amounts of target language data (MENG, M1, M10, M25, M50, and M100) and the monolingual models (MSLO, MT Baseline, EXT Baseline, PG). The last two rows represent the statistics of reference Slovene and English summaries. We cannot compute the perplexity of the EXT model as this purely extractive model outputs human-written sentences.

The English monolingual model MENG generates more than twice as many character as the other models and on average 4 sentences while the other models generate 2 to 3. These numbers are the result of learning, as the dataset of English summaries contains on average more and longer summaries. The M1 model shows that as little as 1k of additional instances is enough to update the number of extracted sentences.

The metrics ROUGE-1, ROUGE-2, and ROUGE-L show similar relations between the compared models. Surprisingly, the zero-shot transfer model MENG scores higher on ROUGE metrics than M1 and M10. The reason for this is that it extracts more sentences, generates longer summary sentences, and repeats the sentences. Analyzing the results of MENG, we noticed that the model sometimes cannot finish a sentence properly, e.g., it generates good content, but does not stop and just continues to generate words. We speculate that the problem is in special tokens (start of the sentence, end of the sentence, etc.) that capture the grammar of source language. These special tokens may be a hidden problem in the cross-lingual seq2seq model transfer.

We manually inspected the returned summaries to assess their readability. M1 does not show any significant readability improvement over MENG, while M10 shows some improvement. MENG often generates long sentences with redundant and rare words, and inserts punctuation at inappropriate places. On the contrary, M1 generates too short sentences and summaries with many missing words. M10 shows an improvement in sentence selection over M1, and improvement in readability over

both M1 and MENG. Still, most of the sentences are not well-formulated, but the meaning is present in almost all of them.

Models M25 and M50 show interesting properties, considering that they produce scores quite close to the models trained on much larger training sets in the target language (i.e. M100 and MSLO). This indicates utility of cross-lingual transfer which can produce useful models with significantly less data.

The PG model scores highest on ROUGE scores, but its perplexity is not on par even with M10. The reason for this is the pure abstractive nature of this model (other models are hybrid extractive-abstractive models). In the generation phase, the abstractive models are not constraint when choosing the content. The manual inspection shows that the PG model generates summaries with higher readability than MT Baseline but much lower than the cross-lingual models trained on sufficient amounts of data.

M100 (cross-lingual model) and MSLO (trained from scratch) are the best models for the Slovene summarization. These two models use the same amounts of training data. With manual inspection, we were unable to conclude which model is better in terms of readability. However, M100 consistently shows better ROUGE scores: ROUGE-1 is improved for 0.60, ROUGE-2 for 0.19, and ROUGE-L for 0.52. This shows that our cross-lingual approach produces better summaries compared to monolingual models even without additional sentence selection mechanism analyzed in Section 5.3.

5.3 Selection of the final output sentences

As explained in Section 4.6, we use our best cross-lingual summarization mode to generate 64 hypotheses for each of the extracted sentences. The candidate sentences are assessed with four heuristics (ROUGE-L, BERTscore, the internal loss value, and perplexity of the language model) and the best is included in the final summary. Table 3 shows the results.

As the baseline result, we report the scores of our best fine-tuned model M100 (taken from Section 5.2), which uses only the internal loss score to select the final output. All the selection heuristics improve the performance of the baseline model. We tested all combinations of the four selection metrics but report only the best one (in the last row of Table 3).

Selection heuristics	ROUGE-1	ROUGE-2	ROUGE-L
M100 with no additional selection	21,67	6,81	19,16
M100 + Transformer LM	22,53	6,83	19,61
M100 + Multilingual BERTScore	24,87	7,41	21,36
M100 + ROUGE-L	24,88	7,38	21,47
M100 + ROUGE-L & BERTScore	**24,97**	**7,43**	**21,50**

Table 3 Selection of the output sentences from the hypotheses generated with the M100 model.

Initially, we hypothesized that two complementary metrics are needed to select the best hypothesis: one for the content and another for the readability. The last line

of Table 3 shows that this is not the case: the best pair of heuristics consists of both content selection metrics, ROUGE-L and BERTscore. These results may be biased since the reported ROUGE metrics are content-based. The manual comparison of models with two complementary metrics and models with both content-based metrics confirmed that the former produced better readable summaries than the latter, but with lower content accuracy. We can conclude that the selection of output hypotheses significantly improves the quality of the output summaries.

5.4 Human evaluation

The automatic summary evaluation is limited in assessment of actual user needs and expectations [32]. For that reason, we organized a small study with human evaluation of generated summaries. For each full text, we used both the reference summary and the automatically generated candidate in a random order.

The task of referees was to assign the accuracy and readability score of a summary (see Table 4 for the scale of scores). The accuracy represents the amount of overlap between the given facts and the summarized information, and the readability measures fluency and comprehensibility of the summary. In our study, each of the 10 articles (two summaries per text, the generated and the reference) were evaluated by eight referees. Referees included three females and five males aged from 23 to 65, with different degrees of education.

Score	Accuracy	Readability
1	none	incomprehensible
2	little	poor
3	a lot of	acceptable
4	most of	good
5	all	flawless

Table 4 The scales for the accuracy and readability scores of summaries.

We report averages and standard deviations of the assigned scores in Table 5. Surprisingly, the accuracy of the reference summaries is lower than the accuracy of the generated summaries. We identified several reasons that explain this result. First, the reference summaries are actually the first paragraphs of news articles and often contain true facts and information that cannot be verified in the text. Unless misleading and speculative, the generated summaries produce verifiable content. Second, the evaluation scores do not directly measure the content quality of summaries. Following the instructions, participants may assign a high score to a summary that contains true but unimportant and irrelevant information. Third, our hybrid summarization model selects and paraphrases sentences. We assume that sometimes participants are lured into thinking that there is a greater content overlap between the text and the generated summary than between the text and a reference summary. Finally, our study is small and the standard deviation of the answers is considerable, therefore the results may be misleading. As anticipated, the readability score of the reference summaries is much higher than it is for the generated summaries.

Type	Accuracy	Readability
Reference	2,85 (1,24)	4,18 (0,96)
Generated	3,06 (1,18)	3,41 (0,94)

Table 5 Average and standard deviation of human assigned accuracy and readability of reference and generated summaries.

5.5 Comparison with related research

We compare our best summarization model (M100 + ROUGE-L & BERTScore) to other existing summarization models for English (as an upper bound of existing technologies) and Slovene. Table 6 shows the results reported by authors of the listed models. In addition to the standard ROUGE scores, we also provide BERTscore where possible. The reported scores are not directly comparable but give a general picture of the success of the proposed cross-lingual approach.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Zidarn [51] Slovene LSTM	23,77	7,97	23,95	0,679
Our M100 + ROUGE-L & BERTScore	24,97	7,43	21,50	0,679
English Chen and Bansal [9]	40,88	17,80	38,54	\
English Zhang et al. [48] - PEGASUS	44,17	21,47	41,11	\

Table 6 Comparison of our best model with related Slovene model and state-of-the-art English models.

The only other neural summarization model for Slovene was built by Zidarn [51] who used a two-layer LSTM neural network with the attention mechanism, copy mechanism, and beam search. The dataset of this model is the same STA news dataset extracted from Gigafida corpus, but the author uses different train, test, and validation splits. Our model scored higher on ROUGE-1 (1.20 difference) but lower on ROUGE-2 (0.54) and ROUGE-L (2.45). The BERTScore results of both models are identical. Given the existing sources of variation (different subsets of the original data, different splits, and the problematic nature of automatic summary evaluation metrics), we can conclude that both models perform similarly.

Table 7 shows human evaluation of our best model and the best model of Zidarn [51]. For both models, human reported scores of the generated and reference summaries are presented. Both models produce acceptable readability scores, but in terms of accuracy, it seems that our model generates more accurate content.

Model	Text type	Accuracy	Readability
Our M100 + ROUGE-L & BERTScore	Reference	2,85 (1,24)	4,18 (0,96)
Zidarn [51] Slovene LSTM	Reference	2,61 (1,39)	3,48 (1,04)
Our M100 + ROUGE-L & BERTScore	Generated	3,06 (1,18)	3,41 (0,94)
Zidarn [51] Slovene LSTM	Generated	1,95 (1,24)	3,10 (1,27)

Table 7 Average and standard deviation of human assigned accuracy and readability of reference and generated summaries.

As the bottom part of Table 6 shows, neither cross-lingual nor monolingual Slovene models can compare to English models in terms of performance. English models are usually trained either on the 4 million instances of the Gigaword dataset, appropriate for headline generation, or the 290k CNN/Daily Mail dataset, which is similar but larger than our Slovene dataset. The English model used in our experiments [9] achieves scores that are almost twice as high compared to our Slovene model. Its results are less misleading and mostly represents facts and information accurately. Many manually inspected summaries show that it omits less important dependent clauses. In our model, this behaviour is less frequent.

PEGASUS [48] is currently one of the best abstractive summarization models. It is based on the transformer neural architecture and presents an interesting novel insight: models are fine-tuned faster and more successfully if they are pretrained on tasks similar to the final task. Authors thus propose two pretraining objectives. One is the BERT masked language model known from [13]. Another is the gap sentence generation that selects and masks whole sentences from documents, and concatenates the gap-sentences into a pseudo-summary. The model is pretrained on two very large corpora. The C4 dataset consists of texts from 350M web-pages (750GB). The HugeNews dataset is even larger with 1,5B articles (3,8TB). The model achieved state of the art performance on 12 summarization tasks.

5.6 Manual analysis of strengths and weaknesses

In this section, we manually analyze three outputs of different quality from our best model (M100 + ROUGE-L & BERTScore) (Tables 8, 9, and 10). In the tables, ”Slovene reference summary” represents the first paragraph of an article. The most important explanatory factor for the differences in quality seems to be the topic of a document. The model generates satisfactory summaries for texts with political and financial content, which represent the majority of the fine-tuning dataset (STA news). For the comprehensibility sake, we manually translated all the texts from Slovene to English, preserving the problems.

The first example in Table 8 demonstrates a good quality result. The summary is short, contains the essential information expressed with well-formulated sentences, and exhibits a certain level of abstraction. It replaces the phrase ”the croatian news agency hina wrote that ... ” with ”foreign news agencies reported that ... ” and cuts off the supplemntary information that starts with ”announcing that austria would...”. In the second sentence, the phrase ”european council president donald tusk” is omitted for no apparent reason. The sentence uses a pronoun ”they” for a replacement of the phrase ”the austrian news agency apa”, which indicates abstractive qualities.

The second example in Table 9 shows that the model can be misleading and factually inconsistent with the text. The mentioned play will not premiere in Ljubljana but in Maribor. The model speculates that the play will start at 8 p.m, although the text says thursday night. The third example in Table 10 shows that the model correctly identifies winners and loosers, but misrepresents the numbers (and some of the names), which was one of the most frequently observed errors.

Human translation of the original Slovene article
the croatian news agency hina reported that the slovenian government had expressed a negative opinion on the austrian control of the border with slovenia, announcing that austria would extend the control of the internal schengen border with slovenia for another six months. the austrian news agency apa also reported about it. hina also reported that slovenian prime minister marjan šarec will meet with the european council president donald tusk and the european commission president jean - claude juncker in an official visit to brussels in the autumn. the latter has recently been the subject of much criticism in ljubljana for its alleged bias in the arbitration dispute between slovenia and croatia. hina wrote that slovenian foreign minister miro cerar, currently visiting washington, expects relations between slovenia and the united states to improve. he intends to better inform the americans about the arbitration dispute between slovenia and croatia, as in his opinion us is not sufficiently acquainted with this problem. the serbian news agency tanjug reported that the slovenian police unions (the slovenian police union and the union of slovenian police), will resume strike activities on monday, which froze in march. tanjug also reported that the serbian president aleksander vučić received today the slovenian ambassador to serbia, vladimir gasparič, on a farewell visit. on this occasion, gasparic expressed his belief that the planning of the visit, which vucic and slovenian president borut pahor had recently discussed, was an additional incentive for good cooperation between the two countries.
Human translation of the Slovene reference summary
foreign news agencies wrote that the slovenian government had expressed a negative opinion on austrian control of the border with slovenia, announcing that austria would extend control of the internal schengen border with slovenia. they also reported that the slovenian police unions would resume preparations for strike activities.
Human translation of the generated summary from Slovene, ROUGE-L = 51,46
foreign news agencies reported that the slovenian government had expressed a negative opinion on austrian control of the border with slovenia. they also reported that slovenian prime minister marjan šarec will meet with european commission president jean - claude juncker on an official visit to brussels in the autumn.
The generated summary in Slovene
tuje tiskovne agencije so poročale o tem , da je slovenska vlada izrazila negativno mnenje o avstrijskem nadzoru na meji s slovenijo . poročale so tudi , da se bo slovenski premier marjan šarec na jesenskem uradnem obisku v bruslju srečal s predsednikom evropske komisije jean - claudom junckerjem .

Table 8 The first example (good quality) of a summary produced by the best cross-lingual summarizer.

6 Conclusion and further work

We developed a neural cross-lingual approach to abstractive summarization. Our solution is based on the pretrained model in the resource-rich language (English), whose outputs are fine-tuned to the target language (Slovene) and further refined with sentence selection heuristics. We first showed that zero-shot transfer is unsatisfactory due to its output following the grammar of the source language. In few-shot transfer, we tested how different amounts of training data in target language used in fine-tuning affects the model and discovered that even small amounts of data in the target language significantly improve the quality of produced summaries. Nevertheless, the quantity and quality of the training sets play a huge role, and the target language dataset (Slovene) is not competitive in either respect. This is most evident when analyzing diverse topics from the Slovene dataset, where better-represented topics are better summarized compared to less represented ones. In addition to the automatic evaluation, we manually analyzed the quality of the results and also conducted a small-scale human evaluation. The assessments show that the accuracy and readability of the generated summaries are acceptable. Two additional contributions of our work are the first Slovene summarization dataset consisting of news articles, and publicly

Human translation of the original Slovene article
the magnificent play of shadows and sound is in the hands of animators barbara jamšek and elene volpi. the show, which will premiere on thursday night, is based on a motif by dennis haseleye's picture book about a pirate trying to catch the moon and with songs from the žmavc press stage. "the story is about a greedy pirate who wants to steal the whole world, and in the end reaches for the moon," director tin grabnar told the news conference today. such a story, in his opinion, is an excellent starting point for a shadow theater performance, where the material world in the form of puppets and other props is placed in relation to the immaterial in the form of light and shadows. the performance takes place on a ship with two sails, set in a recently restored church, and serves both as a stage and a grandstand for spectators. the viewer is placed at the center of the action and, in the words of the author of the artistic image darko erdelja, has the feeling that he is at sea, "limited by matter, but by craving for more". the main language of the play is shadows, not words, as the text of the picture book has been severely curtailed in order to achieve a greater contrast between the material and the immaterial, according to the playwright katarina klančnik kocutar. "appropriating everything material, but at the same time wanting more, even immaterial," is the main theme of the story with characters who have always stirred the human imagination, such as the moon, the sea, pirates. the latter are not only a symbol of greed for material things, but, according to history, also of people on the fringes of society, persecuted for various reasons. due to the absence of lyrics, music, authored by iztok drabik jug, who also used the electric guitar, plays an important role in creating the atmosphere. actresses and animators are barbara jamšek and elena volpi. for them, in addition to learning about the game of shadows and the use of lights, it was a great challenge to play on all sides, as they are surrounded by the audience in the show. since this is not a classic shadow theater performance where the animators are hidden behind a screen, there is a lot of emphasis on the choreography and movement. they had some problems with the acoustics in creating the show, as the church, which otherwise borders the puppet theater and was renovated last year with european funds, lacks technical equipment. according to the director of the mojca theater, they rarely looked for such an ambient performance in order to be able to take advantage of the givens of a sacral building and at the same time test the working conditions in it. otherwise, they are still waiting for the municipal tender to fill their space with a new content.
Human translation of the Slovene reference summary
march brings to the maribor puppet theater the premiere of the play pirate and the moon, directed by tina grabnar. a shadow theater devoted to the relationship between the material and the immaterial was placed in a minorite church, with the church nave serving as a vessel.
Human translation of the generated summary from Slovene, ROUGE-L = 9,30
in the play theater ljubljana) the dennis haseleye's play about a pirate, which is based on a haseleye picture book, will premiere at 8 pm
The generated summary in Slovene
v predstava teatru ljubljana) bodo drevi ob 20. uri premierno uprizorili predstavo dennisa haseleyeja o piratu , ki je nastala po motivih slikanice haseleyeja

Table 9 The second example (misleading) of a summary produced by the best cross-lingual summarizer.

available character-based transformer neural language model. The source code of our system is freely available[3].

The model can be improved in several ways. The quality of the cross-lingual alignment between Slovene and English embeddings is lower than for some other language pairs and could be improved with additional anchor points, such as bilingual dictionary. Recently introduced contextual embeddings such as BERT [13] or ELMo [36] have improved many tasks where they were applied. It would be worth testing their ability in a generative cross-lingual task such as cross-lingual summarization. Further, it may be necessary to increase the vocabulary size because of the rich Slovene morphology. Instead of ROUGE reward, RL step could maximize BERTScore reward. Instead of the two used readability measures (the internal loss function and Slovene language model) used in the selection of the generated sum-

[3] https://github.com/azagsam/cross-lingual-summarization

Human translation of the original Slovene article
formis: desy vahen 14, sešel 5, bračko, pintarič 20, stavbar, polanec, geratič, gajser 15, petranović 2, štern 2. kema puconci: praprotnik flisar 6, vidonja 12, vinkovič, franko 8, škraban 12, banfi 2, donša 4, jerič, koler, frumen 9, sakovič 17, kadiš 2. the volleyball players of the formis suffered another defeat, and the Kome puconci were looking forward to a new victory. rogožanke with the point won did not take a step towards the middle of the scale, so prekmurje with two points remained in its upper half. the introductory set passed in a draw, and in the playoffs the guests showed a more mature game than the hosts and took the lead with 1: 0. also in the second set, no team gained a significant advantage, rogožanke improved their game in attack and defense, and after the lead of 24:23 and the result of 25:25, they won two more points and equalized the result in the sets. the third set was again won by prekmurje, who took advantage of the poor initial blows of the hosts, and in the final they were concentrated enough not to allow a turn. the home team started the fourth set very summery and led all the time. with the result of 16:14, they dominated the field, scored points as if on a conveyor belt and tied the score at 2: 2. the decisive set was started much better by the guests, who took the lead with 5: 1 and 8: 5, then the hosts restored the balance on the field, and with the result of 9: 9, due to errors in reception and attack, three points and also win.
Human translation of the Slovene reference summary
the volleyball players of kema puconci defeated the home team with 3: 2 (21, - 25, 21, - 16, 15) in the match of the 7th round 1 .dol for women in hoče. * sports hall in Hoče, 130 spectators, judges: valentar (straight) and štumfelj (mežica).
Human translation of the generated summary from Slovene, ROUGE-L = 40,00
volleyball players keme puconci beat formis in the 3rd round 1 dol for women with 1: 0 (1: 0 * sports hall, spectators 250, judges: bračko (kranj, štern volleyball players kema puconci are in the 2nd round 1
The generated summary in Slovene
odbojkarice keme puconci so v 3. krogu 1 dol za ženske v gosteh premagale formis z 1:0 (1:0 * športna dvorana , gledalcev 250 , sodnika : bračko (kranj , štern odbojkarice kema puconci so v 2. krogu 1

Table 10 The third example (misrepresented numbers) of a summary produced by the best cross-lingual summarizer.

maries, we could use the recently introduced supervised or unsupervised multilingual readability approach of Martinc et al. [26]. We could improve the quality of the fine-tuning dataset by procuring news articles with the original summary-text splits (instead of the currently used heuristics). Additionally, we could denoise the Slovene dataset by calculating BERTScore scores between reference summaries (i.e. leads) and news article text and retain only the best-matching pairs.

Future studies could investigate how to improve metrics for the abstractive text summarization. One idea is to combine the content-based metrics (ROUGE, BERTScore) with the perplexity measure to ensure both accuracy and readability in the same metric. An interesting problem for future work is how to attain greater levels of abstraction. In cross-lingual and model transfer research, the influence of special tokens should be studied.

Acknowledgements The research was supported by the Slovene Research Agency through research core funding no. P6-0411 and project no. J6-2581. The research was financially supported by European social fund and Republic of Slovenia, Ministry of Education, Science and Sport through projects Quality of Slovene textbooks (KaUč) and Ministry of Culture of Republic of Slovenia through project Development of Slovene in Digital Environment (RSDO). This paper is supported by European Union's Horizon 2020 Programme project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media, grant no. 825153).

References

1. Adams O, Makarucha A, Neubig G, Bird S, Cohn T (2017) Cross-lingual word embeddings for low-resource language modeling. In: Proceedings of the 15th Conference of the European Chapter of the ACL: Volume 1, Long Papers, pp 937–947
2. Aksenov D, Schneider JM, Bourgonje P, Schwarzenberg R, Hennig L, Rehm G (2020) Abstractive text summarization based on language model conditioning and locality modeling. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp 6680–6689
3. Artetxe M, Schwenk H (2019) Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics 7:597–610
4. Baevski A, Auli M (2018) Adaptive input representations for neural language modeling. In: International Conference on Learning Representations. ICLR
5. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations, ICLR
6. Bois R, Leveling J, Goeuriot L, Jones GJ, Kelly L (2014) Porting a summarizer to the French language. In: Proceedings of TALN 2014 (Volume 2: Short Papers), pp 550–555
7. Bojanowski P, Grave É, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5:135–146
8. Chelba C, Mikolov T, Schuster M, Ge Q, Brants T, Koehn P, Robinson T (2014) One billion word benchmark for measuring progress in statistical language modeling. In: Fifteenth Annual Conference of the International Speech Communication Association
9. Chen YC, Bansal M (2018) Fast abstractive summarization with reinforce-selected sentence rewriting. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers, pp 675–686
10. Chi Z, Dong L, Wei F, Wang W, Mao XL, Huang H (2020) Cross-lingual natural language generation via pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence
11. Cohan A, Dernoncourt F, Kim DS, Bui T, Kim S, Chang W, Goharian N (2018) A discourse-aware attention model for abstractive summarization of long documents. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp 615–621
12. Dai Z, Yang Z, Yang Y, Carbonell JG, Le Q, Salakhutdinov R (2019) Transformer-XL: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 2978–2988
13. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Compu-

tational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 4171–4186
14. Dou ZY, Liu P, Hayashi H, Jiang Z, Neubig G (2020) GSum: A general framework for guided neural abstractive summarization. arXiv preprint arXiv:201008014
15. Fecht P, Blank S, Zorn HP (2019) Sequential transfer learning in NLP for German text summarization. In: Proceedings of the 4th edition of the Swiss Text Analytics Conference
16. Gambhir M, Gupta V (2017) Recent automatic text summarization techniques: a survey. Artificial Intelligence Review 47(1):1–66
17. Graff D, Kong J, Chen K, Maeda K (2003) English Gigaword. Linguistic Data Consortium, Philadelphia 4(1):34
18. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. In: Language Resources and Evaluation Conference
19. Grusky M, Naaman M, Artzi Y (2018) Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp 708–719
20. Hu B, Chen Q, Zhu F (2015) LCSTS: A large scale Chinese short text summarization dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp 1967–1972
21. Krek S, Arhar-Holdt Š, Erjavec T, Čibej J, Repar A, Gantar P, Ljubešić N, Kosem I, Dobrovoljc K (2020) Gigafida 2.0: The reference corpus of written standard Slovene. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp 3340–3345
22. Kryściński W, Rajani N, Agarwal D, Xiong C, Radev D (2021) BookSum: A collection of datasets for long-form narrative summarization. ArXiv preprint 2105.08209
23. Lample G, Conneau A, Ranzato M, Denoyer L, Jégou H (2018) Word translation without parallel data. In: International Conference on Learning Representations, ICLR
24. Li L, Forăscu C, El-Haj M, Giannakopoulos G (2013) Multi-document multilingual summarization corpus preparation, part 1: Arabic, English, Greek, Chinese, Romanian. In: Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pp 1–12
25. Lin CY, Hovy E (2002) Manual and automatic evaluation of summaries. In: Proceedings of the ACL-02 Workshop on Automatic Summarization, Volume 4, pp 45–51
26. Martinc M, Pollak S, Robnik-Šikonja M (2021) Supervised and unsupervised neural approaches to text readability. Computational Linguistics pp 1–39
27. Merrouni ZA, Frikh B, Ouhbi B (2019) Automatic keyphrase extraction: a survey and trends. Journal of Intelligent Information Systems 54:391–424
28. Mihalcea R (2004) Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp 170–173

29. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:13013781
30. Mikolov T, Le QV, Sutskever I (2013) Exploiting similarities among languages for machine translation. arXiv:13094168
31. Nallapati R, Zhou B, dos Santos C, Gulcehre C, Xiang B (2016) Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning, pp 280–290
32. Novikova J, Dušek O, Curry AC, Rieser V (2017) Why we need new evaluation metrics for NLG. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 2241–2252
33. Ouyang J, Song B, McKeown K (2019) A robust abstractive system for cross-lingual summarization. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 2025–2031
34. Over P, Dang H, Harman D (2007) DUC in context. Information Processing & Management 43(6):1506–1520
35. Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp 1532–1543
36. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of NAACL-HLT, pp 2227–2237
37. Qi W, Yan Y, Gong Y, Liu D, Duan N, Chen J, Zhang R, Zhou M (2020) ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp 2401–2410
38. Ruder S, Vulić I, Søgaard A (2019) A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research 65:569–631
39. Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp 379–389
40. Scialom T, Dray PA, Lamprier S, Piwowarski B, Staiano J (2020) MLSUM: The multilingual summarization corpus. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp 8051–8067
41. See A, Liu PJ, Manning CD (2017) Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1073–1083
42. Straka M, Mediankin N, Kocmi T, Žabokrtskỳ Z, Hudeček V, Hajic J (2018) SumeCzech: Large Czech news-based summarization dataset. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC
43. Suppa M, Adamec J (2020) A summarization dataset of Slovak news articles. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp 6725–6730

44. Tu Z, Lu Z, Liu Y, Liu X, Li H (2016) Modeling coverage for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 76–85
45. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp 6000–6010
46. Vaswani A, Bengio S, Brevdo E, Chollet F, Gomez A, Gouws S, Jones L, Kaiser Ł, Kalchbrenner N, Parmar N, et al. (2018) Tensor2Tensor for neural machine translation. In: Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pp 193–199
47. Vinyals O, Fortunato M, Jaitly N (2015) Pointer networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2, pp 2692–2700
48. Zhang J, Zhao Y, Saleh M, Liu P (2020) Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning, PMLR, pp 11328–11339
49. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2019) BERTScore: Evaluating text generation with BERT. arXiv:190409675
50. Zhu J, Wang Q, Wang Y, Zhou Y, Zhang J, Wang S, Zong C (2019) NCLS: Neural cross-lingual summarization. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 3045–3055
51. Zidarn R (2020) Automatic text summarization of Slovene texts using deep neural networks. University of Ljubljana, Faculty of computer and information science, Ljubljana, (MSc thesis in Slovene)

Appendix B: TeMoTopic Temporal Mosaic Visualization of Topic Distribution, Keywords, and Context

TeMoTopic: Temporal Mosaic Visualization of Topic Distribution, Keywords, and Context

Shane Sheehan, Saturnino Luz
University of Edinburgh
United Kingdom
shane.sheehan@ed.ac.uk
s.luz@ed.ac.uk

Masood Masoodian
Aalto University
Finland
masood.masoodian@aalto.fi

Abstract

In this paper we present *TeMoTopic*, a visualization component for temporal exploration of topics in text corpora. *TeMoTopic* uses the temporal mosaic metaphor to present topics as a timeline of stacked bars along with related keywords for each topic. The visualization serves as an overview of the temporal distribution of topics, along with the keyword contents of the topics, which collectively support detail-on-demand interactions with the source text of the corpora. Through these interactions and the use of keyword highlighting, the content related to each topic and its change over time can be explored.

1 Introduction

Many text corpora, such as news articles, are temporal in nature, with the individual documents distributed across a span of time. As the size and availability of text corpora have continued to increase in recent years, effective analysis of the content of corpora has become challenging. Taking the temporal nature of most corpora into account when analysing the text makes it more difficult to describe the corpora and to interpret intuitively the results of analysis.

Topic modeling techniques, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), have been used to automatically generate topic groups in text corpora. These topics can help in understanding the contents of a corpus by using keywords and topic association probabilities generated by the topic modelling technique. However, interpreting the results of the techniques is not always easy, and the results can seem counter-intuitive when looking only at the weighted keyword lists. Therefore, visualization techniques have been used extensively to help with the interpretation of the large number of topics generated by these models. The same is true of temporal topic modeling techniques, such as Dynamic Topic Modeling (Blei and Lafferty, 2006), which require additional visualization techniques to aid intuitive understanding of the temporal segmentation of the topics and their related keywords.

In this paper, we propose *TeMoTopic* as a contribution to the collection of visualization techniques for exploring the temporal distribution of topics in text corpora through the use of temporal mosaics. *TeMoTopic* adopts a space-filling approach to show topic distribution over time, and presents keywords related to each topic at the overview level of the visualization. The visualization is interactive and, in contrast to many other techniques, enables direct investigation of the source documents associated with individual topics and keywords. This allows the user to get a general sense of the meaning of a topic through its associated keywords, as well as providing the ability to dive into the details of the related documents.

2 Related Work

2.1 Temporal Topic Visualization

Topic visualization systems are an active research area, with a variety of approaches for visualizing different aspects of topic model outputs, topic hierarchies, and topic evolution. In this paper, we only focus on related work in the area of temporal topic evolution and topic visualization of text corpora. While some methods address the temporal structuring of topics in short texts in the context of meetings and dialogues (Luz and Masoodian, 2005; Sheehan et al., 2019), in recent years, visualization of temporal topic evolution for larger text collections has been based on flow diagrams. An early example of such an approach is *ThemeRiver* (Havre et al., 2002), with later additions such as *TextFlow* (Cui et al., 2011), *TopicFlow* (Malik et al., 2013), *ThemeDelta* (Gad et al., 2015) and *RoseRiver* (Cui et al., 2014).

While *TeMoTopic* and flow-based temporal topic visualizations are similar, we expect they could

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 56–61
April 19, 2021 © Association for Computational Linguistics

Task	Description
Visualize Topics	Visualize topic in terms of extracted keywords
Overview of Document - Topic Relations	View documents related to a topic
Remove Topics from the visualization	Topic removal from overview
Filtering Documents	View a subset of documents for a topic
Perform Set Operations	Enable exclusion/inclusion of documents in the corpus
Show and Cluster Similar Topics	Enable identification of similar topics
Perform Cluster Operations	Enable grouping of similar topics
Annotating Topics	Allow for labelling of the topics
Visualize Topic Change	View topic distribution and keywords over time

Table 1: Visualization tasks for topic model exploration.

form complementary components used in model assessment tools that are used to evaluate model quality. Flow diagrams are, for instance, useful for getting a high-level overview of many topics across long spans of text. *TeMoTopic*, on the other hand, aims to provide support for detailed viewing of a subset of topics and shorter timeslices, which are not possible in a flow diagram. As such, we envisage that other existing visualization tools which include a flow diagram component – such as *LDAExplore* (Ganesan et al., 2015), *VISTopic* (Yang et al., 2017), *ParallelTopics* (Dou et al., 2011) and *TIARA* (Wei et al., 2010) – could be further expanded to include a temporal mosaic visualization, in the style of *TeMoTopic*. The largest benefit to this integration would come from enabling intuitive interactive filtering of the source documents based on the temporal topic and keyword distribution.

2.2 Topic Visualization Tasks

The design of a visualization tool should clearly be motivated by concrete tasks relevant to the end-users of the intended tool. Munzner's *nested model for visualization design and validation* (Munzner, 2009) describes steps that can be taken to mitigate threats to the validity of a visualization design. The first of the four levels of this design model is the characterization of domain specific tasks which should be supported by the visual encoding.

Ganesan et al. (2015) identify key tasks, in the design description of *LDAExplore*, which should be supported by visualizations that aim to help users explore the results of Latent Dirichlet Allocation (LDA). Since LDA is one of the most commonly used topic modelling techniques for text corpora, these key tasks could be generalized to other techniques where a corpus is also split into topics, and keywords associated with those topics are extracted.

In addition, Ganesan et al. (2015) argue that the results of LDA can be counter-intuitive, and that the ability to explore and interact with the document set should make the topic and word distributions more intuitive and insightful. Table 1 shows the eight tasks identified by Ganesan et al. (2015), as well as one additional task which we consider to be important for visualizing temporal topics. The table also includes a brief description of the tasks which are fully described by Ganesan et al. (2015).

Theses tasks describe a need for topic overview with document detail available on-demand, this follows the well-known visual information seeking mantra proposed by Shneiderman (1996). Interactions around viewing, filtering, removing, and combining topics and documents should also be supported. Finally, we include an additional task for visualizing topic changes over time. This modifies the *Visualize Topics* task, such that the change in distribution and keywords across is available to explore.

3 TeMoTopic: Temporal Mosaic Topic visualization

Figure 1 shows the *TeMoTopic* visualization tool. It consists of two juxtaposed views (Javed and Elmqvist, 2012): the temporal mosaic (left), and the document view (right). The design of the temporal mosaic is based on a visualization proposed by Luz and Masoodian (2007), and further expanded in our previous temporal mosaic visualizations *TeMoCo* visualization (Sheehan et al., 2019) and *TeMoCo-Doc* visualization (Sheehan et al., 2020), which have been used to link transcripts of meetings to document reports in a medical context.

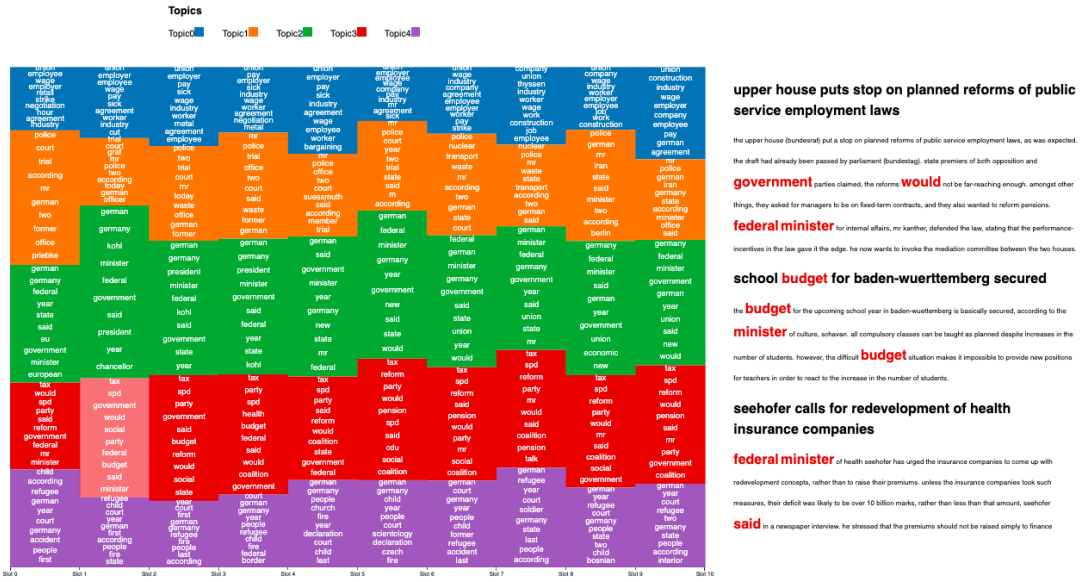


Figure 1: *TeMoTopic* visualization, showing the temporal mosaic view (left) and the document view (right), showing the selected keywords for the red topic in the second timeslice (red tile on the bottom left).

Figure 2: *TeMoTopic* visualization, showing the temporal mosaic view (left) and the filtered document view (right), with the word "german" selected from a temporal topic timeslice (orange tile on the top left).

3.1 Prototype

The temporal mosaic encoding was designed using Mackinlay's ranking (Mackinlay, 1986) of visual variables (Bertin, 1983), such that the visualization uses a perceptually efficient static encoding of the key data attributes. Horizontal position is used to emphasize the temporal order of the topics, and topic distribution per timeslice is encoded using vertical length. Each tile in the mosaic represents a single combination of topic and timeslice. The height of each tile represents its topic weight in that timeslice.

The top ten keywords which describe the associated temporal topic are placed within the tile, and can be scaled to encode the keyword topic probability, using area in a manner similar to keyword scaling in text visualizations such as word clouds (Viegas et al., 2009). Although the keywords are currently presented in order of descending topic probability, in future work alternative keyword presentation styles such as alphabetized lists and word clouds will be compared in terms of their effectiveness for comparison between the tiles. The categorical topics are encoded using color, allowing topics weights and keyword changes to be examined across the span of timeslices.

The mosaic visualization provides an overview of the topic distribution and associated keywords over time. However, as the number of topics and

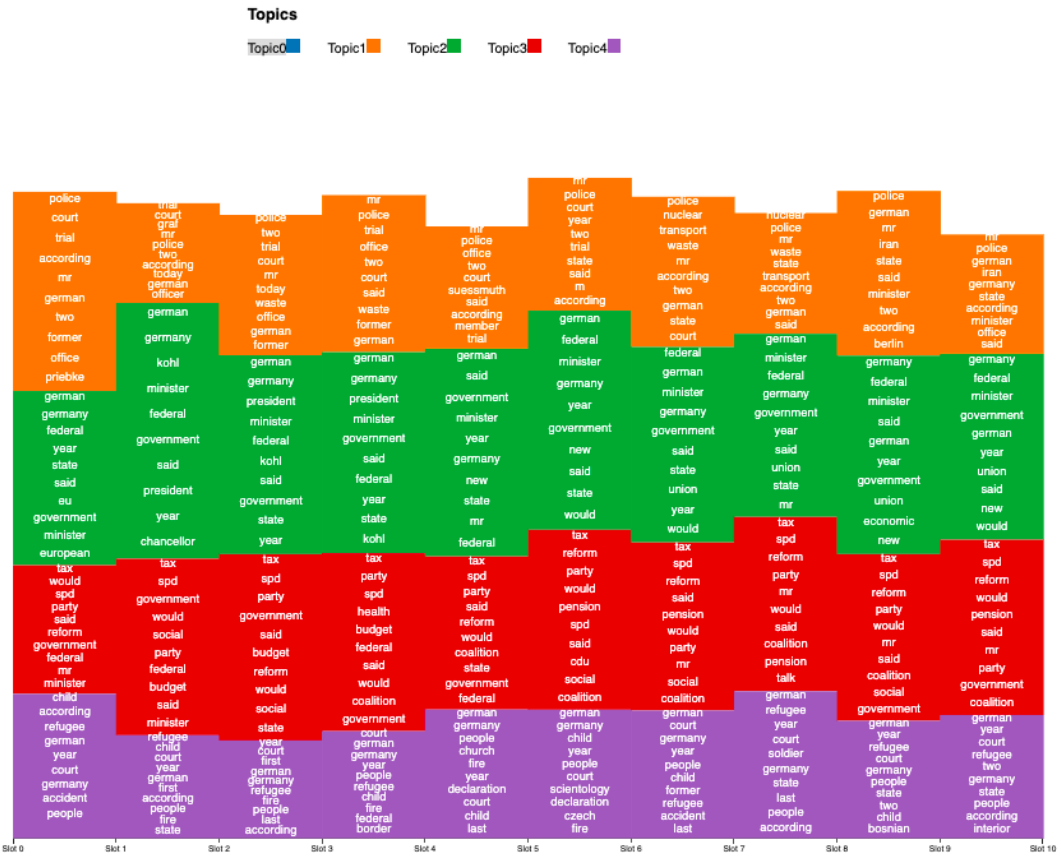


Figure 3: *TeMoTopic* filtered temporal mosaic view after the blue topic was selected for removal via clicking on the legend.

timeslices increase, if the visualization area is kept at a fixed size, the overview would become more abstract, cluttered, and difficult to examine for individual tiles and keywords. To maintain readability, the visualization can extend both horizontally and vertically to accommodate more topics and timeslices. The user can pan and zoom to get the detailed views of topics and keywords, or a higher-level view of the entire temporal topic space. The removal interaction is particularly useful when the number of topics is large, since filtering out topics that are not relevant to the current analysis allows for more of the detail to be presented on a single screen.

The temporal mosaic, as currently described, addresses two of the tasks from Table 1, namely *Visualize Topics* and *Visualize Topic Change*. To facilitate *Overview of Document - Topic Relations*, the document view (Figure 1, right) was created and linked, via click interactions, to the temporal mosaic (Figure 1, left). The document view is used to display the documents associated with a temporal topic tile. When a coloured tile is selected in the temporal mosaic, the related articles are presented in a scroll box and, the keywords from the topic tile are highlighted in the text. If keyword weights (or probabilities) are provided, the highlighted words are scaled accordingly. This dual combination of views and described interactions, support the user in investigating the meaning of a topic, and by investigating the differences between the topic timeslices, temporal document similarities and differences can be revealed.

Although it is useful to view the entirety of a topic, *Filtering Documents* is a task that was also identified as important to facilitate. One simple and intuitive way to do this with the temporal mosaic is by clicking on individual keywords rather than on the entire topic tile. This will cause the document view to display only documents from the related topic timeslice which contain the selected keywords, as shown in Figure 2. Selection from

multiple topics is also possible, and the keywords are highlighted in the related topic colour to differentiate between topics.

The final interaction supported by this version of *TeMoTopic* is the removal of topics from the temporal mosaic. To do this, a topic can be selected from the legend shown above the temporal mosaic (Figure 3, top). Alternatively right-clicking on a topic removes all the other topics except the selected one. In the example shown in Figure 3, the blue topic has been removed from the temporal mosaic. When topics are removed, the temporal mosaic no longer fills the entire vertical space of the visualization. This interaction is useful when dealing with a large number of topics of which only a few are of interest for the analysis.

3.2 Implementation

The visualization tool[1] is implemented as a single-page web application using the *D3.js* framework (Bostock et al., 2011) . It takes two JavaScript Object Notation (JSON) files as input: the first file contains topic, keyword, timeslice, weights, and associated filenames, and the second input file is simply a JSON structure containing the documents with filename used as the retrieval key. Sample Python scripts are provided for generating topics and keywords on the sample dataset and for preparing the visualization input files from the model output.

The current version of *TeMoTopic* was designed to be model agnostic, and can even be used for tasks unrelated to topic model exploration. For example, metadata attributes such as the source of the news articles or their author could be used in place of topics. Keywords could be extracted using any available technique, including simple frequency lists. The visualization could also be used for corpus comparison and even cross-lingual analysis using entire corpora as replacements for the topics.

However, in our implementation we make use of dynamic topic modelling (Blei and Lafferty, 2006) to identify temporal topics and keywords in a subset of the *de-news*[2] corpus of German-English parallel news. The dataset consists of transcribed German radio broadcasts which were manually translated into English. Between 1996 and 2000 volunteers selected and transcribed five to ten of these news broadcasts per day and added them to the dataset. In the examples of *TeMoTopic*, shown in Figures 2, 1 and 3, we selected a ten month span of the dataset and presented the four largest topics. The choice of time span and topic number was only for presentation and to exemplify the interface features. We did not attempt to choose a time period or number of topics based on prior knowledge of the news relevant at the time in Germany. We present our examples to describe the interface and interactions, rather than as an analysis of the dataset, and we choose to draw no conclusions about the dataset contents and topics.

[1]The software and working example are available at https://github.com/sfermoy/TeMoCo.

[2]http://homepages.inf.ed.ac.uk/pkoehn/publications/de-news/

4 Conclusions

While many other temporal visualization techniques, such as ThemeRiver (Havre et al., 2002), offer some of the functionality for temporal visualization of topics or visualization of content changes, they do not feature implicit linking between the visualization and the underlying content documents. We consider this to be the main contribution of *TeMoTopic* visualization and its distinguishing feature with regards to the state of the art. As such, determining the necessity and validity of this approach in the identified domain is an important step before further development of the visualization prototype. Future work will, therefore, include evaluating the usability of a future iteration of the system with domain experts in both news analysis and topic modelling.

Acknowledgments

The work of the first and second authors is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

References

Jacques Bertin. 1983. *Semiology of Graphics*. University of Wisconsin Press.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. 2011. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421.

W. Cui, S. Liu, Z. Wu, and H. Wei. 2014. How hierarchical topics evolve in large text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2281–2290.

Wenwen Dou, Xiaoyu Wang, Remco Chang, and William Ribarsky. 2011. Paralleltopics: A probabilistic approach to exploring document collections. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pages 231–240. IEEE.

Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, Tom Ewing, Keith N Hampton, and Naren Ramakrishnan. 2015. Themedelta: Dynamic segmentations over temporal topic models. *IEEE transactions on visualization and computer graphics*, 21(5):672–685.

Ashwinkumar Ganesan, Kiante Brantley, Shimei Pan, and Jian Chen. 2015. Ldaexplore: Visualizing topic models generated using latent dirichlet allocation. *arXiv preprint arXiv:1507.06593*.

S. Havre, E. Hetzler, P. Whitney, and L. Nowell. 2002. Themeriver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20.

Waqas Javed and Niklas Elmqvist. 2012. Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, pages 1–8.

Saturnino Luz and Masood Masoodian. 2005. A model for meeting content storage and retrieval. In *Proceedings of the 11th International Multimedia Modelling Conference*, MMM '05, pages 392–398.

Saturnino Luz and Masood Masoodian. 2007. Visualisation of parallel data streams with temporal mosaics. In *Proceedings of the 11th International Conference Information Visualization*, IV '07, pages 197–202.

Jock Mackinlay. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141.

Sana Malik, Alison Smith, Timothy Hawes, Panagis Papadatos, Jianyu Li, Cody Dunne, and Ben Shneiderman. 2013. Topicflow: Visualizing topic alignment of twitter data over time. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 720–726.

T. Munzner. 2009. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928.

Shane Sheehan, Pierre Albert, Masood Masoodian, and Saturnino Luz. 2019. Temoco: A visualization tool for temporal analysis of multi-party dialogues in clinical settings. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 690–695. IEEE.

Shane Sheehan, Saturnino Luz, Pierre Albert, and Masood Masoodian. 2020. Temoco-doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA. Association for Computing Machinery.

Ben Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages*, pages 336–343.

Fernanda B. Viegas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144.

Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162.

Yi Yang, Quanming Yao, and Huamin Qu. 2017. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics*, 1(1):40–47.