Compressive approaches for cross-language multi-document summarization

Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, Andréa Carneiro Linhares

Data & Knowledge Engineering 125 (2020) 101763

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Compressive approaches for cross-language multi-document summarization

Elvys Linhares Pontes [a,*], Stéphane Huet [a], Juan-Manuel Torres-Moreno [a,b], Andréa Carneiro Linhares [c]

[a] Laboratoire Informatique d'Avignon, Avignon Université, 339 Chemin des Meinajariès, Avignon, 84140, France
[b] Département GIGL, Polytechnique Montréal, C.P. 6079, succ. Centre-ville, Montréal (Québec) H3C 3A7, Canada
[c] Curso de Engenharia da Computação, Universidade Federal do Ceará, Rua Coronel Estanislau Frota, 563, Sobral-Ceará, CEP 62.010-560, Brazil

ARTICLE INFO

Keywords:
Cross-language text summarization
Sentence compression
Multi-sentence compression
Optimization

ABSTRACT

The popularization of social networks and digital documents has quickly increased the multilingual information available on the Internet. However, this huge amount of data cannot be analyzed manually. This paper deals with Cross-Language Text Summarization (CLTS) that produces a summary in a different language from the source documents. We describe three compressive CLTS approaches that analyze the text in the source and target languages to compute the relevance of sentences. Our systems compress sentences at two levels: clusters of similar sentences are compressed using a multi-sentence compression (MSC) method and single sentences are compressed using a Neural Network model. The version of our approach using multi-sentence compression generated more informative French-to-English cross-lingual summaries than extractive state-of-the-art systems. Moreover, these cross-lingual summaries have a grammatical quality similar to extractive approaches.

1. Introduction

Nowadays, most books and newspapers have digital or audio versions while the popularization of social networks (such as Facebook, Twitter, YouTube, among others) and news websites have enabled a great increase in the amount of data trafficked over the Internet on a wide variety of subjects. Readers, besides not having the time to go through this amount of information, are not interested in all the proposed subjects and generally select the content of their interest. Another limiting factor is the language of messages, a lot of news being available in languages that readers do not know or have little knowledge of.

Cross-Language Text Summarization (CLTS) consists in analyzing a document in a language source to get its meaning and, then, generate a short, informative and correct summary of this document in a target language. The methods developed for CLTS can be classified, like the Text Summarization (TS) field, depending on whether they are extractive, compressive or abstractive [1–3]. The extractive TS produces a summary by concatenating the most relevant sentences of the documents; the compressive TS generates a summary by removing non-relevant information of sentences; lastly, the abstractive TS produces a summary with new sentences that are not necessarily contained in the source documents.

Many of the state-of-the-art methods for CLTS are of the extractive class. They mainly differ on how they compute sentence similarities and alleviate the risk that translation errors are introduced in the produced summary. Recent systems have used compressive and abstractive approaches [4–6] to improve the informativeness and the grammatical quality of summaries. However,

* Corresponding author.
E-mail addresses: elvys.linhares-pontes@alumni.univ-avignon.fr (E.L. Pontes), stephane.huet@univ-avignon.fr (S. Huet).

https://doi.org/10.1016/j.datak.2019.101763
Received 19 January 2019; Received in revised form 25 October 2019; Accepted 1 November 2019
Available online 5 November 2019

these approaches require specific resources for each language [5] or external data and combination of different methods [4,6] that limit the adaptability of these methods to generate summaries into other languages.

Inspired by compressive TS methods in monolingual analysis [2,7–12], we combine sentence and multi-sentence compression methods for the CLTS problem to generate more informative cross-lingual summaries. This article extends our previous work [13] and its contribution is threefold. (i) We improve our previous model for Sentence Compression (SC) that compresses sentences by removing non-relevant words. In line with the works [10,14], we add an attention mechanism to our previous Long Short Term Memory (LSTM) model in order to identify the gist of sentences. (ii) We adapt our Multi-Sentence Compression (MSC) approach [15] to compress small clusters of similar sentences. We simplified our Integer Linear Programming (ILP) formulation in order to optimize the complexity and focus only on the cohesion of words and a list of keywords.[1] In particular, we deleted the analysis of 3-grams from the objective function used in our previous work [15]. This simplification improved the analysis of the relevance between coherence of words (grammaticality) and the keywords (informativeness) on the generation of compressions. It also made our model faster to find the best path in the word graph. (iii) We carry out both automatic and manual evaluations to compare the quality of our compressive cross-lingual approach with state-of-the-art extractive methods. Our system using MSC not only generates more informative cross-lingual summaries but also achieves grammatical scores similar to extractive cross-lingual summaries. Unfortunately, adding the attention mechanism to the sentence compression approach was not enough to generate fluent compressions for long and complex source sentences. The low performance of the SC approach also degraded the quality of the SC+MSC compressions.

The rest of the paper is organized as follows. In the next section, we review previous work in CLTS. In Section 3, we detail our compressive CLTS approaches using sentence and multi-sentence compression methods. The experimental results in the context of a French-to-English multi-document summarization task are analyzed in Section 4 and we make our conclusions in Section 5.

2. Related work

The first studies in cross-language document summarization analyzed the information in only one language [16,17]. Two typical CLTS schemes are the early and the late translations. The first scheme translates the source documents into the target language, then it summarizes the translated documents using only information of the translated sentences. The late translation scheme does the reverse: it first summarizes the documents using abstractive, compressive or extractive methods, then it translates the summary into the target language.

Leuski et al. [16] proposed an early translation method to generate English headlines for Hindi documents. Orasan and Chiorean [17] implemented the late translation approach; they produced summaries with the Maximal Marginal Relevance (MMR) method from Romanian news articles and then automatically translated the summaries into English.

Recent methods have improved the quality of cross-language summarization using a translation quality score [4,18,19] and the information of the documents in the source and target languages [5,20]. These methods are described in the next two subsections.

2.1. Machine translation quality

Machine translation evaluation aims to assess the correctness and quality of the translation. Usually, a human reference translation is provided, and various methods and metrics have been developed for comparing the system-translated text and the human reference text.

Another possibility is the use of automatic methods to estimate translation quality (see for example the quality estimation shared task of the WMT conference [21]). The translation quality of a sentence can be estimated at word-level, phrase-level and sentence-level. The estimation at word-level aims to detect errors for each token in Machine Translation (MT) outputs by deciding if a token is correct in the translation. An incorrect word can cause several errors in the translation, especially in its local context. The estimation at phrase-level is similar to the word-level, i.e. the estimation verifies whether a phrase is correct in the translation. Finally, the estimation of translation quality at sentence-level aims to generate scores for the translations according to post-editing effort, i.e. the percentage of needed edits, post-editing time, and so on.

Wan et al. [18] proposed a framework to generate English-to-Chinese CLTS based on the translation quality of sentences. They used basic features (such as sentence length, sub-sentence number, percentage of nouns and adjectives) and parse features (such as depth, number of noun phrases and verbal phrases in the parse tree) to train a support vector machine regression method to predict the translation quality of a pair of English–Chinese sentences. Their dataset is composed of 1736 pairs of English–Chinese sentences with translation quality scores in a range from 1 to 5 (1 represents "very bad" and 5 corresponds to "excellent" translations). The salience of English sentences is estimated based on the translation quality and informativeness scores in order to select informative English sentences with a high translation quality:

$$\text{score}(s_i) = \lambda \cdot \text{translation\_score}(s_i) + (1 - \lambda) \cdot \text{informative\_score}(s_i) \quad (1)$$

where $\text{translation\_score}(s_i)$ and $\text{informative\_score}(s_i)$ are the translation quality prediction and informativeness score of the sentence s_i, respectively; and $\lambda \in [0, 1]$ is a parameter controlling the influence of the two factors. Finally, they translated the most relevant English sentences to produce the Chinese summary.

[1] In this work, we consider keywords as the significant words that define the topic of a document.

Similarly to Wan et al. [18], Boudin et al. [19] developed an ϵ-Support Vector Regression (SVR) based on the automatic NIST metrics as an indicator of quality to predict the translation quality score. They automatically translated English documents into French using Google Translate, then they estimated the translation quality of a sentence based on different features (sentence length, number of punctuation marks, perplexities of source and target sentences using different language models, etc.). They incorporated the translation quality score in the PageRank algorithm [22] to calculate the relevance of sentences based on the similarity between the sentences and to perform English-to-French cross-language summarization (Eqs. (2)–(4)):

$$\mathrm{p}(v_i) = (1-d) + d \times \sum_{v_j \in \mathrm{pred}(v_i)} \frac{\mathrm{score}(s_i, s_j)}{\sum_{v_k \in \mathrm{succ}(v_i)} \mathrm{score}(s_k, s_i)} \mathrm{p}(v_i) \tag{2}$$

$$\mathrm{score}(s_i, s_j) = \mathrm{similarity}(s_i, s_j) \times \mathrm{prediction}(s_i) \tag{3}$$

$$\mathrm{similarity}(s_i, s_j) = \frac{\sum_{w \in s_i, s_j} \mathrm{freq}(w, s_i) + \mathrm{freq}(w, s_j)}{\log(|s_i|) + \log(|s_j|)} \tag{4}$$

where d is the damping factor, $\mathrm{pred}(v_i)$ and $\mathrm{succ}(v_i)$ are the predecessor and successor vertices of the vertex v_i, $\mathrm{prediction}(s_i)$ is the translation quality score of the sentence s_i, and $\mathrm{freq}(w, s_i)$ is the frequency of the word w in the sentence s_i.

Inspired by the phrase-based translation models, Yao et al. [4] presented a phrase-based model to simultaneously perform sentence scoring, extraction and compression of sentences. They developed a CLTS framework based on a submodular term of compressed sentences and a bounded distortion penalty term to estimate the quality of the translation. Their summary scoring $F(sum)$ measure was defined over a summary sum as:

$$\mathrm{F}(sum) = \sum_{p \in sum} \sum_{i=1}^{\mathrm{count}(p, sum)} d^{i-1} \mathrm{g}(p) + \sum_{s \in sum} \mathrm{bg}(s) + \eta \sum_{s \in sum} \mathrm{dist}(\mathrm{pbd}(s)) \tag{5}$$

where d is a constant damping factor to penalize repeated occurrences of the same phrases, $\mathrm{count}(p, sum)$ is the number of occurrences of the phrase p in the summary sum and η is the distortion parameter for penalizing the distance between neighboring phrases in the derivation. Finally, $\mathrm{g}(p)$ is the frequency of p in the document, $\mathrm{bg}(s)$ is the bigram score of sentence s, $\mathrm{pbd}(s)$ is the phrase-based derivation of the sentence s and $\mathrm{dist}(.)$ is the distortion penalty term based on the reordering probability of the phrase-based translation models.

2.2. Joint analysis of source and target languages

Wan [20] proposed to leverage both the information in the source and in the target language for cross-language summarization. In particular, he presented two graph-based TS methods (SimFusion and CoRank) for the task of English-to-Chinese CLTS. The first method linearly fuses the English-side and Chinese-side similarities for measuring Chinese sentence relevance. In a nutshell, this method adapts the PageRank algorithm to calculate the relevance of sentences, where the weight arcs are obtained by the linear combination of the cosine similarity[2] of pairs of sentences for each language:

$$\mathrm{relevance}(s_i^{tr}) = \mu \sum_{j \in D, j \neq i} \mathrm{relevance}(s_j^{tr}) \cdot \tilde{C}_{ji}^{tr} + \frac{1-\mu}{n} \tag{6}$$

$$C_{ij}^{tr} = \lambda \cdot \mathrm{cosine}(s_i^{tr}, s_j^{tr}) + (1-\lambda) \cdot \mathrm{cosine}(s_i^{sr}, s_j^{sr}) \tag{7}$$

where s_i^{sr} and s_i^{tr} represent the sentence i of a document D in the source and target languages, respectively, μ is a damping factor, n is the number of sentences in the document and $\lambda \in [0, 1]$ is a parameter to control the relative contributions of the two similarity values. C^{tr} is normalized to $\tilde{C}^{tr}$ to make the sum of each row equal to 1.

The CoRank method adopts a co-ranking algorithm to simultaneously rank both source and target sentences by incorporating mutual influences between them. It considers a sentence as relevant if this sentence is heavily linked with other sentences in each language separately (source–source and target–target language similarities) and between languages (source–target language similarity).

Recently, Wan et al. [6] carried out the cross-language document summarization task by extraction and compression through the ranking of multiple summaries in the target language. They analyzed many candidate summaries in order to produce a high-quality summary for every kind of documents. These candidate summaries were generated using multiple text summarization and machine translation methods, e.g. bilingual submodular function, multiple machine translations and multi-sentence compressions. Their method used a top-K ensemble ranking based on features at several levels and perspectives (word-level, sentence-level, summary-level, readability-related and source-side features) that characterized the quality of a candidate summary.

By contrast with Wan et al. [6] who generated extractive and compressive CLTS, Zhang et al. [5] analyzed Predicate-Argument Structures (PAS) to get an abstractive English-to-Chinese CLTS. They built a pool of bilingual concepts and facts represented by the bilingual elements of the source-side PAS and their target-side counterparts from the alignment between source texts and Google Translate translations. They used word alignment, lexical translation probability and 3-g language model to measure the quality and the fluency of the Chinese translation, and the CoRank algorithm [20] to measure the relevance of the facts and concepts in both

[2] The cosine similarity between two vectors u and v associated with two sentences is defined by $\frac{u \cdot v}{\|u\| \ \|v\|}$ in the [0,1] range.

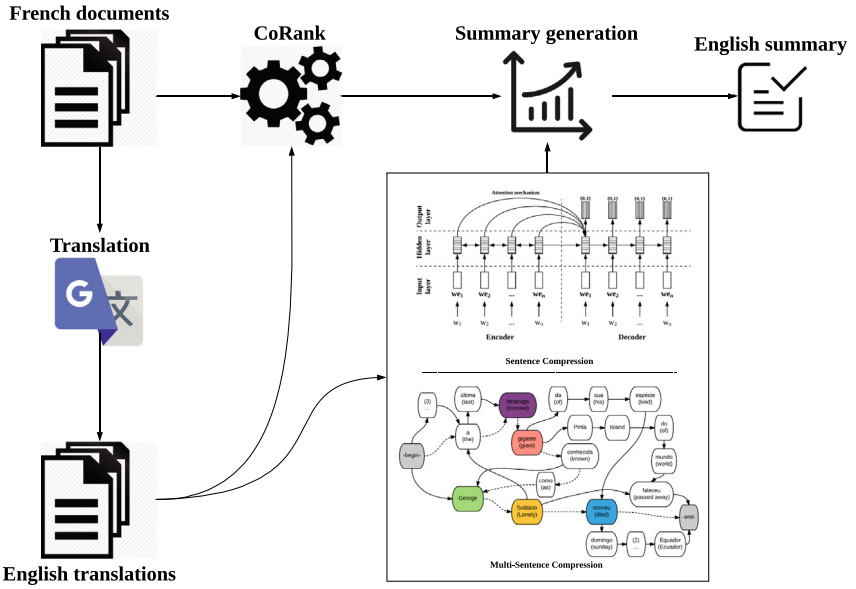

Fig. 1. Overview of our compressive CLTS model to generate French-to-English cross-lingual summaries using SC and MSC methods.

languages. Finally, summaries were produced by fusing bilingual PAS elements with an ILP algorithm to maximize the saliency and the translation quality of the PAS elements. Their ILP model used the pool of bilingual concepts (facts) and their scores to generate summary sentences composed of a concept and at least one core fact.

Among the existing approaches for summarization, abstractive TS methods have a greater capacity to generate summaries more similar to the human abstracts. However, this kind of summarization demands large datasets available in a language to train Neural Network (NN) models. On the contrary, extractive summarization approaches do not require specific resources to generate summaries; nevertheless, these extracted sentences may contain redundant and/or non-relevant information, thus reducing the informativeness of summaries. Finally, some compressive methods only need a few resources in several languages to generate summaries. Therefore, we devised an MSC approach optimized to TS which generates compressions guided by keywords and the cohesion of words. This approach is easily adaptable to other languages and can still improve the informativeness of cross-lingual summaries. Moreover, we also developed an NN method with an attention mechanism to delete non-relevant words and to only retain the main information of source sentences.

3. Our approach

Inspired by the compressive TS methods in monolingual analysis [2,7–12], we adapt sentence and multi-sentence compression methods for the French-to-English CLTS problem to just keep the main information. An LSTM model is built to analyze a sentence and decide which words remain in the compression. We also use an ILP formulation to compress similar sentences while analyzing both grammaticality and informativeness.

Following our previous study [13], we combine sentence and multi-sentence compressions to generate more informative cross-lingual summaries. We expanded this method in two ways.

Multi-document summarization consists of analyzing several documents about the same topic and generating a short and concise summary that describes the main information of these documents. Commonly, the parts relating to the main topic are repeated with complementary information in several documents. Therefore, sentences are grouped in clusters based on their similarity in the source and target languages to get the gist of these documents. We simplified our previous MSC method [13,15] to just focus on the cohesion of words and a list of keywords to guide the compressions with the core information of these clusters. This simplification improved the analysis of the relevance between coherence of words (grammaticality) and the keywords (informativeness) on the generation of compressions. This improvement has also made our model faster than our previous model.

A second extension of the approach relies on compression techniques of a single sentence by deletion of words [10,14]. Still with the idea to generate more informative summaries, we extended our previous SC method by adding an attention mechanism to compress sentences that stand alone during the clustering step required by the MSC step.

The following subsections describe the architecture of our system in detail (Fig. 1).

3.1. Preprocessing

Initially, French texts are translated into English using the Google Translate system, which was used in the majority of the state-of-the-art CLTS methods.

Instead of analyzing separate words in SC and MSC, we consider chunks in order to improve the grammatical quality of compressions. To realize this chunk-level tokenization, we used the Stanford CoreNLP tool and the syntactic pattern $\langle (ADJ)^*(NP|NC)+\rangle$ for the English sentences [23]. This annotator tool, which integrates jMWE[3] [24], detects various expressions, e.g., phrasal verbs ("*look after*"), proper names ("*New York*"), compound nominals ("*credit card*") or idioms ("*kick the bucket*"). For instance, the sentence "The giant tortoise known as Lonesome George died Sunday at the Galapagos National Park in Ecuador" will be tokenized as "'The' 'giant tortoise' 'known as' 'Lonesome George' 'died' 'Sunday' 'at' 'the' 'Galapagos National Park' 'in' 'Ecuador'".

Finally, sentences are clustered according to their similarities, the sentences with a similarity score higher than a threshold θ remaining in the same group. The similarity score of a pair of sentences i and j is defined by the cosine similarity in the source and target languages:

$$\text{sim}(i,j) = \sqrt{\text{cosine}(s_i^{sr}, s_j^{sr}) \times \text{cosine}(s_i^{tr}, s_j^{tr})} \tag{8}$$

where s_i^{sr} and s_i^{tr} represent a sentence i in the source and target languages, respectively.

3.2. Relevance of sentences

In order to analyze the information in the source and target languages, our approach incorporates the CoRank method [20] that ranks sentences in both languages based on their mutual influences. CoRank scores sentences based on their similarities in each language separately, but also between languages (Eqs. (9)–(13)).

$$\mathbf{u} = \alpha \cdot (\tilde{\mathbf{M}}^{\mathbf{tr}})^T \mathbf{u} + \beta \cdot (\tilde{\mathbf{M}}^{\mathbf{srtr}})^T \mathbf{v} \tag{9}$$

$$\mathbf{v} = \alpha \cdot (\tilde{\mathbf{M}}^{\mathbf{sr}})^T \mathbf{v} + \beta \cdot (\tilde{\mathbf{M}}^{\mathbf{srtr}})^T \mathbf{u} \tag{10}$$

$$M_{ij}^{\text{sr}} = \begin{cases} \text{cosine}(s_i^{sr}, s_j^{sr}), & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$M_{ij}^{\text{tr}} = \begin{cases} \text{cosine}(s_i^{tr}, s_j^{tr}), & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$\text{M}_{ij}^{\text{srtr}} = \sqrt{\text{cosine}(s_i^{tr}, s_j^{tr}) \times \text{cosine}(s_i^{sr}, s_j^{sr})} \tag{13}$$

where $\mathbf{M^{sr}}$ and $\mathbf{M^{tr}}$ are normalized to $\tilde{\mathbf{M}}^{\mathbf{sr}}$ and $\tilde{\mathbf{M}}^{\mathbf{tr}}$, respectively, to make the sum of each row equal to 1. $\mathbf{u}$ and $\mathbf{v}$ denote the saliency scores of the target and source language sentences, respectively; α and β specify the relative contributions to the final saliency scores from the information in the same language and the information in the other language, with $\alpha + \beta = 1$.

3.3. Sentence and multi-sentence compression

In order to only retain the main information, we apply sentence and multi-sentence compressions to improve the informativeness of English sentences and, consequently, of cross-lingual summaries.

3.3.1. Sentence compression

The Sentence Compression (SC) problem is here seen as the task to delete non-relevant words in a sentence [9,10,14,25]. In a similar way to [14], our method extends the LSTM model described in [10] to compress a sentence by deletion of words. In few words, our model follows a sequence-to-sequence paradigm using an LSTM model with an attention mechanism to verify which words of a sentence s remain in the compression (Fig. 2). The words in a sentence c are represented by their word embeddings. Then, a first LSTM encodes this sentence [26] and a second LSTM with attention selects the sequence of words that are kept in the compression. The attention mechanism decides which input region of the encoder is focused on in order to generate the next output [27].

LSTM with the attention mechanism has an input i_t and a memory state m_t that are updated at time step t (Eqs. (14)–(23)).

$$x_t = [we_t, c_t] \tag{14}$$

$$i_t = \text{sigm}(W_1 x_t + W_2 h_{t-1}) \tag{15}$$

$$i'_t = \tanh(W_3 x_t + W_4 h_{t-1}) \tag{16}$$

$$f_t = \text{sigm}(W_5 x_t + W_6 h_{t-1}) \tag{17}$$

$$o_t = \text{sigm}(W_7 x_t + W_8 h_{t-1}) \tag{18}$$

$$m_t = m_{t-1} \odot f_t + i_t \odot i'_t \tag{19}$$

$$h_t = m_t \odot o_t \tag{20}$$

[3] http://projects.csail.mit.edu/jmwe/.

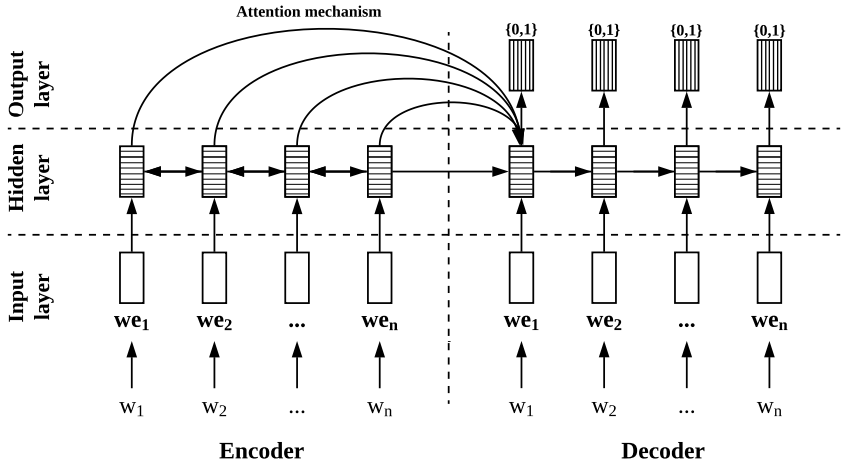

Fig. 2. The words are represented by the word embedding representations in the input layer. The attention mechanism improves decoding. The output layer is composed of 0 (remove) or 1 (keep).

where the operator $\odot$ denotes element-wise multiplication, we_t is the word embedding of the word at the time step t, c_t is the context vector, $sigm$ is the sigmoid function, the matrices $W_1, \ldots, W_8$ and the vector h are the parameters of the model, and all the non-linearities are computed element-wise. The context vector c_t at time t is calculated as a sum of all hidden states of the encoder weight:

$$c_t = \sum_{j=1}^{T} \alpha_{tj} \cdot h_E^j \tag{21}$$

$$r_{tj} = v_\alpha^T \tanh(W_\alpha h_{t-1} + U_\alpha h_E^j) \tag{22}$$

$$\alpha_{tj} = \text{softmax}(r_{tj}) \tag{23}$$

where the probability α_{tj} represents the importance of each hidden state of the encoder h_E^j in the prediction of the current state h_t. Differently from [10,14], we analyze the sentence at the chunk level, so we remove a chunk only if all words of this chunk were deleted in the SC process described above.

3.3.2. Multi-sentence compression

For the clusters that have more than a sentence, we use a Chunk Graph (CG) to represent them [13,15] and an ILP method to compress these sentences in a single, short, and hopefully correct and informative sentence. In this case, the example "'The giant tortoise known as Lonesome George died Sunday at the Galapagos National Park in Ecuador" will be represented as "'the' 'giant_tortoise' 'known_as' 'Lonesome_George' 'died' 'Sunday' 'at' 'the' 'Galapagos_National_Park' 'in' 'Ecuador'".' Among several state-of-the-art MSC methods [11,12,15,28], our system incorporates our previous ILP model for the MSC [13,15] to generate compressions of clusters of similar sentences using multi-word chunks (Chunk Graph).

The construction of the chunk graph is similar to the procedure described by Filippova to create her word graph (more details in [28]). Initially, the CG 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 analyzed and added individually to the CG. For each analyzed 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.

In this work, we simplified our ILP formulation to only analyze the cohesion of chunks and keywords. This formulation enables a better analysis of informativeness (list of keywords) and grammaticality (cohesion of chunks) of CGs with different sizes.

The cohesion of chunks is calculated from the frequency and the position of these chunks in sentences, according to Eqs. (24)–(25):

$$\text{w}(i,j) = \frac{\text{cohesion}(i,j)}{\text{freq}(i) \times \text{freq}(j)}, \tag{24}$$

$$\text{cohesion}(i,j) = \frac{\text{freq}(i) + \text{freq}(j)}{\sum_{s \in D} \text{diff}(s,i,j)^{-1}}, \tag{25}$$

where $\text{freq}(i)$ is the chunk frequency mapped to the vertex i and the function $\text{diff}(s,i,j)$ refers to the distance between the offset positions of chunks i and j in the sentence s of the cluster of similar sentences D containing these two chunks.

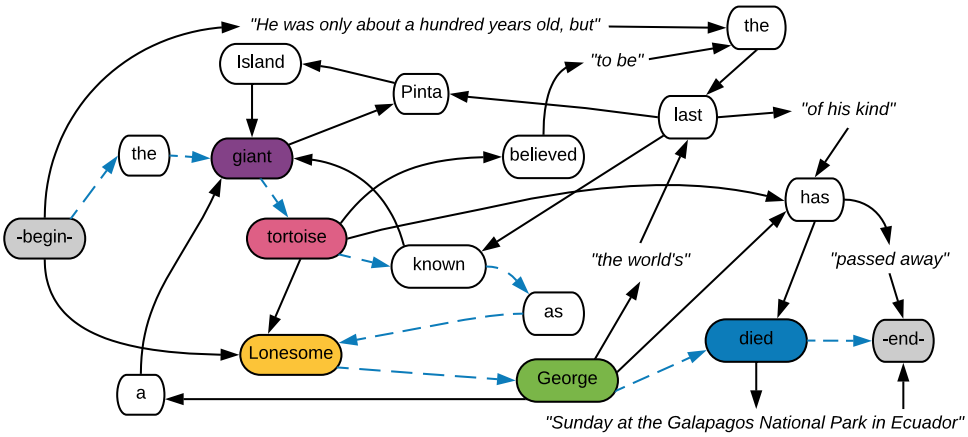

Fig. 3. A word graph model with labels of a cluster of similar sentences. The dotted line represents the compression proposed by our approach.

In order to generate compressions with the gist of documents, we consider keywords at the global (all texts of a topic) and local (cluster of similar sentences) levels to have the essence of both the documents and the cluster of similar sentences. LDA is a topic model that generates topics based on word frequency from a set of documents [29]. This model provides the most significant words that define a topic. Therefore, we consider these significant words as keywords for this topic. In the context of multi-document summarization, we consider that clusters of similar sentences (local level) only describe one topic since it is made of semantically close sentences related to a specific news item. In this case, LDA looks for words in each cluster that represent better each topic. Then, we consider these words as keywords of this cluster. On the global level, we are interested in knowing the most important subjects discussed in all documents. In order to accomplish this, we consider all documents have only a topic and LDA identifies the most relevant words that represent this topic.

Our ILP formulation looks for a path in CG that is composed of chunks with a good cohesion between them and with a maximum number of keywords:

$$\text{Minimize} \left(\sum_{(i,j)\in A} \text{w}(i,j) \cdot x_{i,j} - c \cdot \sum_{k \in K} b_k \right) \tag{26}$$

where x_{ij} indicates the existence of the arc (i, j) in the solution, $\text{w}(i, j)$ is the cohesion of the chunks i and j Eq. (24), 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.[4] Eq. (26) is a simplified version of our previous work [15]. This simplified version does not contain the relevance score, the binary variables and the constraints for the 3-gram analysis that makes the system faster. However, the complexity of the model still NP-Hard.

We calculate the 50 best solutions according to the objective Eq. (26) having at least eight words and at least one verb. Specifically, we find the best solution, then we add a constraint in the model to avoid this solution and repeat this process 50 times to find the other solutions.

The optimized score Eq. (26) explicitly takes into account the size of the generated sentence. Contrary to Filippova's method, sentences may have a negative score because we subtract from the cohesion value of the path the introduced scores for keywords. Therefore, we use the exponential function to ensure a score greater than zero. Finally, we select the sentence with the lowest final score Eq. (27) as the best compression.

$$\text{score}_{norm}(s) = \frac{e^{\text{score}_{opt}(s)}}{\|s\|}, \tag{27}$$

where $\text{score}_{opt}(s)$ is the score of the sentence s from Eq. (26). For more details, see [15].

Fig. 3 illustrates the word graph with labels (colored nodes) of the following sentences:

1. Lonesome George, the world's last Pinta Island giant tortoise, has passed away.
2. The giant tortoise known as Lonesome George died Sunday at the Galapagos National Park in Ecuador.
3. He was only about a hundred years old, but the last known giant Pinta tortoise, Lonesome George, has passed away.
4. Lonesome George, a giant tortoise believed to be the last of his kind, has died.

In this example, our approach generated a path between $-begin-$ and $-end-$ with all labels in order to maximize the informativeness. This path represents the compression "*The giant tortoise known as Lonesome George died.*" that is grammatically correct and contains the main information of the cluster.

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

3.4. Summary generation

The last step of summarization is the generation of summaries. This step selects the sentences with the most relevant CoRank scores. Then, these sentences are replaced by their compressions if these compressions are available. Finally, the summary is generated by concatenating these sentences/compressions, while sentences/compressions redundant with the ones that have already been selected are put aside.

4. Experimental results

The most recent systems [4–6] are not publicly available and are not easily replicable. They resort to specific resources that limit their approaches to a pair of languages. Moreover, deep learning techniques may help to generate more informative summaries; however, these models need large training data to analyze specific language features. Unlike these works, our approaches use compression methods that are easily adapted to several languages to improve the informativeness of sentences. For a fair comparison, we looked for systems that also are easily adaptable to other languages. Despite its simplicity, CoRank shows to be a strong baseline by outperforming other approaches and being language-independent [6].

For these reasons, we chose to compare the performance of our approach with extractive state-of-the-art methods (the early translation, the late translation, the SimFusion and the CoRank methods [20]). Following the idea presented by Wan [20], the early and late translations are based on the SimFusion method, the differences between the systems being on the similarity metrics (Eq. (7)) computed either in the target language (early translation) or in the source language (late translation) [20]. SimFusion and CoRank methods use $\lambda = 0.75$[5] and $\alpha = \beta = 0.5$, respectively. Our approach, named Compressive CLTS (CCLTS), is available in three versions: SC, MSC and SC+MSC. The first version compresses all sentences with the SC method; in the second version, clusters of similar sentences are compressed using the MSC method and the rest of the sentences are extracted; and the last version applies both MSC to clusters of similar sentences and SC to other sentences.

Only sentences with more than 15 words were compressed and compressions with more than 10 words were preserved to avoid short outputs with little information. The MSC method selects the 10 most relevant keywords per topic and the 3 most relevant keywords per cluster of similar sentences to guide the MSC compression generation. All systems produce summaries composed of 250 words with the most relevant sentences, while the redundant sentences are discarded. We apply the cosine similarity measure with a threshold θ of 0.5 to create clusters of similar sentences for MSC and to remove redundant sentences in the summary generation.[6]

Our LSTM model just has one layer with 256-dimensional embeddings and it uses a pre-trained word embeddings[7] with 300-dimensional embeddings. This model was trained on the publicly released set of 200,000 sentence-compression pairs.[8]

4.1. Dataset

We performed the tests using the MultiLing Pilot 2011 dataset [31]. This dataset contains several language versions (Arabic, Czech, French, Greek, Hebrew and Hindi languages). Each language version is composed of 10 topics, each topic having 10 source texts and 3 reference summaries. Each reference summary contains a maximum of 250 words. Specifically, we used the French source documents and the reference summaries in English in this work.

4.2. Automatic and manual evaluations

The most important features of CLTS are informativeness and grammaticality. Informativeness measures how informational is the generated text. As references are assumed to contain the key information, we calculated informativeness scores counting the n-grams in common between the system output and the reference summaries. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) measure developed by Lin [32] compares the differences between the distribution of words of the candidate summary and a set of reference summaries. The comparison is made splitting into n-grams both the candidate and the reference to calculate their intersection. Standard n-gram values for ROUGE are 1-gram and 2-gram, both expressed as:

$$\text{ROUGE} - n = \frac{\sum_{n-grams} \in \{Sum_{can} \cap Sum_{ref}\}}{\sum_{n-grams} \in Sum_{ref}}, \tag{28}$$

where n is the n-gram order, Sum_{can} the candidate summary and Sum_{ref} the reference summary. A third common ROUGE-n variation is ROUGE-SUγ. This ROUGE-n variation takes into account skip units (SU) $\leq \gamma$.

Considering the limitations of the automatic evaluation to analyze the grammaticality and the informativeness of cross-lingual summaries, three annotators manually evaluated the cross-lingual summaries in two aspects: grammaticality and informativeness. The informativeness is rated with scores from 1 (summaries without relevant information) to 5 (summaries with the main information of source documents). The grammaticality also has the same range; summaries with several errors have a score of 1 and summaries without grammatical errors has a score of 5.

[5] SimFusion achieved better results with $\lambda = 0.75$ for Chinese-to-English CLTS in [20].

[6] In a preliminary study [30], we analyzed the similarity of clusters of similar sentences on three MSC corpora in different languages (French, Portuguese and Spanish). Their levels of similarity are 0.46, 0.51, and 0.47; therefore, we defined the threshold of similarity as 0.50 to group similar sentence in the same cluster.

[7] Publicly available at: code.google.com/p/word2vec.

[8] github.com/google-research-datasets/sentence-compression/tree/master/data.

Table 1
ROUGE F-scores for the French-to-English CLTS using the MultiLing Pilot 2011 dataset.

Methods	ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline.early	0.4165	0.1021	0.1607
Baseline.late	0.4142	0.1023	0.1589
SimFusion	0.4173	0.1035	0.1606
CoRank	0.4623*	0.1321	0.1926*
CCLTS.SC	0.4352	0.1259	0.1809
CCLTS.MSC	**0.4743***	**0.1369**	**0.1947***
CCLTS.SC+MSC	0.4517	0.1311	0.1852

*Indicates that the results are statistically better than the SimFusion and baselines with 95% confidence interval.

Table 2
Manual evaluation scores for the French-to-English CLTS using the MultiLing Pilot 2011 dataset.

Methods	Informativeness		Grammaticality	
	Average	Std. Dev.	Average	Std. Dev.
Baseline.early	2.9	0.8	3.9	0.5
Baseline.late	2.8	0.7	4.0	0.5
SimFusion	2.9	0.7	4.0	0.5
CoRank	3.3	0.4	**4.3**	**0.7**
CCLTS.SC	3.2	0.6	3.7	0.7
CCLTS.MSC	**3.5**	**0.4**	4.1	0.7
CCLTS.SC+MSC	3.1	0.7	3.4	1.0

Table 3
Reference summary.

Two years after the seizure of Royal Navy personnel by Iran, two inquiries, that examined the British Ministry's of Defence handling, identified weaknesses in training, communications and the handling of intelligence as well as "collective failure of judgement". The fifteen sailors and marines, from the frigate HMS Cornwall, were captured by Iranian border guards on March 23 in the Persian Gulf, while they were inspecting, in accordance with UN Security Council Resolution 1723, a ship believed to be smuggling cars into Iraq. The UK insisted they were operating in Iraqi waters, while Iran claimed they entered illegally into Iran's territorial waters and that they could face charges of espionage. If those charges were brought against them, the result would be heavy punishment by current Iranian law. On 28 March, British Prime Minister froze all bilateral business deals with Iran. The next day, Iran announced that it will "suspend" the releasing of 15 British personnel, due to the political ballyhoo by London. The EU called the Iranian seizure a "clear breach" of international law. Meanwhile, footage of all 15 British personnel had been broadcast on Iranian TV, with one of the sailors saying that the soldiers were in Iranian waters at the time of their detainment. The British government claimed that the confessions were extracted under duress. Few days later, Iranian President announced that he would free them as a "gift to the British people". The fifteen British navy personnel landed at Heathrow on 5 April, after thirteen days of captivity.

4.3. Results

The results of the automatic evaluation using the MultiLing Pilot 2011 dataset are described in Table 1. The baselines, especially the late translation approach, had the worst scores. The joint analysis of the source and target languages outperformed the SimFusion and other baselines. Analyzing the versions of our approach highlighted that the SC model removed relevant information of sentences, thus achieving lower ROUGE scores than CoRank. CCLTS.MSC generated more informative summaries and led to the best ROUGE scores. Finally, the SC+MSC version achieved better results than other systems but still did not reach the highest ROUGE scores measured when using MSC alone.

Table 2 shows the manual evaluation of cross-lingual summaries. All versions of our system ratified the good results of automatic evaluation and generated more informative summaries than the early, the late and the SimFusion. Our system using MSC obtained the highest score for informativeness. As regards the grammaticality, CoRank generated more grammatical compressions but our MSC approach achieved scores similar to other extractive baselines, which proves the generation of compressive summaries with a good grammaticality. Our SC method removed too much relevant information, which reduced the informativeness and the grammaticality of compressions. Section 4.3.1 provides the analysis of informativeness and grammaticality of an example using the CoRank and all versions of our approach, then we discuss these evaluations in Section 4.3.2.

4.3.1. Example analysis

We carried out the analysis of an example of French-to-English CLTS extracted from the Multilingual Pilot 2011 dataset. Table 3 shows a reference summary of the cluster of source documents that describes the capture of fifteen sailors and marines by Iranian border guards. Tables 4–7 show the cross-lingual summaries generated by the CoRank and the three versions of our system. The extractive cross-lingual summary generated by the CoRank method is presented in Table 4. Even using an extractive approach, this summary contains some grammatical errors.

Table 4
Cross-lingual summary generated by the CoRank method.

On Thursday, British Prime Minister Tony Blair said in a television interview that if the 15 sailors and soldiers who were arrested by the Iranian forces were not released, then Britain would be forced to "enter a new phase". operations, and that Iran only has a few days to free the 15 soldiers and sailors. According to Reuters, the United Kingdom sent a 15-page preliminary statement to the UN Security Council "deploring" the continued arrest and support for the British position that soldiers were operating in Iraqi waters as members of the United Nations Security Council. the Iraqi Multinational Force under the mandate of the Security Council ... and at the request of the Iraqi government. The Iranian National Security Council has announced it will "suspend" the release of 15 British sailors and soldiers arrested by Iranian forces on March 23. The defense minister said Royal Navy sailors were "engaged in routine boarding operations in Iraqi territorial waters" and completed their inspection of the suspect ship when they were surrounded by Iranian forces. He added that he had "asked Mr. Blair not to prosecute these 15 soldiers because they confessed their penetration of Iranian territorial waters," apparently implying that the British military were on a secret mission in Iranian waters, and should not have confessed to the television being in there.

Table 5
Cross-lingual summary generated by the CCLTS.SC method.

The British government has asked for the release of the military. the forces were in Iranian waters, and continues. First-hand information on the capture and detention by Iran of the 15 Royal crew British similar to the two that were seized by Iran on March 23, 2007. Errors identified in the response to Iran's capture of Royal Navy soldiers. Iranian media said the British sailors had "shouted for joy" at the news. The Iranian government initially located the incident in Iraqi waters. Britain says they will not negotiate the release of their soldiers. "HMS Cornwall frigates and soldiers were inspecting, in accordance with UN Security Council Resolution 1723 The president said" [after the meeting] they [were] free. All 27 members of the union agreed on the content of the communiqué. British sailors detained by Iran will be "tried for espionage" Sunday, March 25, 2007. We want to resolve in peace and dialog the disagreements we have with your government.

Table 6
Cross-lingual summary generated by the CCLTS.MSC method.

A video of the 15 sailors and soldiers aired on the iranian forces were in iranian waters when they were arrested. The United Kingdom has frozen all bilateral economic relations with Iran until the 15 British sailors and soldiers arrested by Iranian forces on march 23. The defense minister said Royal Navy sailors were "engaged in routine boarding operations in Iraqi territorial waters" and completed their inspection of the suspect ship when they were surrounded by Iranian forces. He added that he had "asked Mr. Blair not to prosecute these 15 soldiers because they confessed their penetration of Iranian territorial waters," apparently implying that the British military were on a secret mission in Iranian waters, and should not have confessed to the television being in there. *Iranian president Mahmoud Ahmadinejad announced that the 15 British sailors as a gift to the British people.* The United Kingdom is ready to move to "a new phase" if British soldiers and sailors are not released by Iran in the days that follow. Two investigations into the capture of Royal Navy soldiers by Iran in March 2007 determined that it was not the result of "a point of failure or human error of a particular individual, but rather than an unfortunate accumulation of factors" and that it resulted in a "collective error of judgment" by allowing those who were involved to be paid to detail these events in front of the media.

Table 7
Cross-lingual summary generated by the CCLTS.SC+MSC method.

A video of the 15 sailors and soldiers aired on the Iranian forces were in Iranian waters when they were arrested. The United Kingdom has frozen all bilateral economic relations with Iran until the 15 British sailors and soldiers arrested by Iranian forces on march 23. Iranian president Mahmoud Ahmadinejad announced that the 15 British sailors as a gift to the British people. Cornwall frigate were inspecting, in accordance with UN security council resolution 1723. The Australian reported that a website "operated by associates of Mahmoud Ahmadinejad" declared that the 15 British soldiers who had been arrested by the Iranian revolutionary guards could be accused of espionage. The British government has asked for the release of the military. Iran said Tuesday that soldiers and sailors are being treated "humanly" and that they are "in good health". Errors identified in the response to Iran's capture of royal crew. what needs to be done when engaging with people like the Iranian government understand that sanctions can be taken if they are not prepared to be reasonable. The UE reiterates its call for the immediate and unconditional release of British royal navy soldiers. British similar to the two that were seized by Iran on March 23, 2007. united, a boat into Iraq which proved unfounded after inspection when Iranian ships surrounded the sailors.

Our SC method compresses all sentences, by generating short compressions (Table 5). SC method attempts to reproduce the principle of its training dataset that eliminates the words of the first sentences of news to reproduce their title. This procedure normally works well when the source sentence has a direct sentence with straight ideas. However, the source sentences may have complex syntactic structures and different ways to explain the facts, which produces summaries with grammatical errors but also with less relevant information. Let us notice that CCLTS.SC generated shorter summaries because our system removes short sentences/compressions (fewer than 10 words) and redundant sentences/compressions from the summaries.

Table 6 shows the cross-lingual summary for CCLTS.MSC. This summary is composed of three compressions and other sentences are extracted from the source documents. These compressions enabled the generation of summaries with more subjects than CoRank. Unfortunately, the clusters of similar sentences have sizes and levels of similarity between the sentences smaller than the MSC dataset [30,33]. These characteristics may generate summaries with sentences that combine different subjects, which can reduce their concision and readability.

Finally, Table 7 describes the cross-lingual summary of SC+MSC version. The compressions generated by MSC improved the informativeness of SC version; however, SC+MSC summary contains the combination of errors generated by MSC and SC, which reduced the grammatical quality of summaries.

4.3.2. Discussion

The early and late translations achieved poor results with respect to other systems, which proves that the texts in each language provide complementary information to estimate the relevance of sentences. This also establishes that the analysis of sentences in the target language plays a more important role to generate informative French-to-English cross-lingual summaries. As seen before

for English-to-Chinese [20] and French-to-English [13] CLTS, the CoRank method generates better results than the baselines and SimFusion because it considers the information in each language separately and together, while the baselines restrict the analysis of sentence similarity to one language separately and the SimFusion method only analyzes a linear combination of sentence similarities in each language.

It is expected that a piece of information found in several texts is relevant for a topic. In accordance with this principle, the MSC method looks for repeated information and generates a short compression with selected keywords that summarize the main information. The two kinds of keywords (global and local) guide MSC to produce a compression linked to the main topic of the documents and to the specific information presented in the cluster. With respect to CoRank, our MSC version improved the informativeness of summaries by generating shorter sentences with the main information, which enabled the addition of other relevant information in the summaries.

With regard to SC, this compression method eliminated much relevant information in our experiments. This observation may be explained by the small size of the corpus we used to train our NN (200,000 parallel sentence-compression instance), while the system described in [10] could benefit from a corpus of about two million instances. In this case, the SC approach attempts to compress all kinds of sentences (simple and complex grammatical); however, the available training dataset is too small to have enough compression examples of long sentences with several subjects and complex syntactic structure. Besides, this training dataset is not suitable for compressing all kinds of sentences because its parallel sentence-compression instances are composed of first sentences of news articles and their titles. These instances are simpler than compressing other news sentences that have more complex structures. A possible solution to overcome these problems is the use of tree-based and sentence-based SC approaches, such as [34,35], to generate more correct and informative compressions for complex sentences.

Whereas the CCLTS.MSC version leaves unchanged the sentences that do not have similar sentences, the SC+MSC version involves the SC model to compress these sentences. As the CCLTS.SC system has lower performance than the pure extractive CoRank method, the SC+MSC also had lower results than the MSC version.

A difference between the SC and MSC approaches is that MSC uses global and local keywords to guide the compression by preserving the main information, while the SC method does not realize this kind of analysis. The CG model helps the generation of more correct compressions independently of the length and subject of sentences. The SC method compresses first sentences of news that normally describe the main idea of the news in a straight way. However, SC are applied here for all kinds of sentences, e.g. sentences with complex syntactic structure and/or with several subjects. Our approach generates poor results for these kinds of sentences. Another difference between them is that MSC does not need a training corpus to generate compressions.

Previous compressive and abstractive approaches in the state of the art [4,5] improved the informativeness of cross-lingual summaries. However, most of them are limited to a pair of languages because their methods require specific resources or libraries that are available only in a specific language. In a previous work [36], we made a version of our approach using the MSC method to generate compressive cross-lingual summaries from {English, French, Portuguese, Spanish} to {English, French} languages. Despite the differences and the specificities among these several pairs of languages, our approach was more stable than extractive methods and was able to generate more informative cross-language summaries with consistent ROUGE scores measured in several languages.

All in all, the joint analysis of the source and target languages provides a better analysis of documents and helps the generation of more informative cross-lingual summaries. On the one hand, the SC model deletes relevant information, thereby reducing the informativeness of summaries. On the other hand, the MSC method proves to be a good alternative to compress redundant information and to preserve relevant one. Finally, the CCLTS.MSC improved the informativeness of cross-lingual summaries without losing the grammatical quality.

5. Conclusion

Cross-Language Text Summarization (CLTS) produces a summary in a target language from documents written in a source language. It implies a combination of the processes of text summarization and machine translation. Unfortunately, this combination produces errors, thereby reducing the quality of summaries. Joint analysis allows CLTS systems to extract relevant information from source and target languages, which improves the generation of extractive cross-lingual summaries. Recent methods have proposed compressive and abstractive approaches for CLTS; however, these methods use frameworks or tools that are available in only a few languages, limiting the adaptability of these methods to other languages.

Unlike the sentence compression system (CCLATS.SC) that needs a large training dataset to generate compressions of good quality, the multi-sentence compression version of our system (CCLATS.MSC) generates more informative cross-lingual summaries than extractive CLTS systems. Moreover, it has the advantage of not requiring a training corpus to generate summaries of good quality. In addition, our method can be easily adapted for other languages.

There are several avenues worth exploring from this work. We want to extend the attention mechanism in our SC model to take into account keywords in order to guide the SC process. Tree-based SC approaches can be used to compress long and complex sentences in order to mitigate the poor performance of NN approaches for these types of sentences. Finally, we also want to test a semantic sentence similarity approach in CLTS to carry out an analysis of the impact of the semantic analysis in the generation of cross-lingual summaries.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially financed by the European Project CHISTERA-AMIS ANR-15-CHR2-0001.

References

[1] J.-M. Torres-Moreno, Automatic Text Summarization, Wiley and Sons, London, 2014.
[2] C. Li, F. Liu, F. Weng, Y. Liu, Document summarization via guided sentence compression, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2013, pp. 490–500.
[3] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, H. Li, Generative adversarial network for abstractive text summarization, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018.
[4] J.-g. Yao, X. Wan, J. Xiao, Phrase-based compressive cross-language summarization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 118–127.
[5] J. Zhang, Y. Zhou, C. Zong, Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (10) (2016) 1842–1853.
[6] X. Wan, F. Luo, X. Sun, S. Huang, J.-g. Yao, Cross-language document summarization via extraction and ranking of multiple summaries, Knowl. Inf. Syst. (2018).
[7] X. Qian, Y. Liu, Fast joint compression and summarization via graph cuts, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 1492–1502.
[8] C. Li, Y. Liu, F. Liu, L. Zhao, F. Weng, Improving multi-documents summarization by sentence compression based on expanded constituent parse trees, in: EMNLP, ACL, 2014, pp. 691–701.
[9] J. Yao, X. Wan, J. Xiao, Compressive document summarization via sparse optimization, in: Proceedings of the 24th International Conference on Artificial Intelligence, in: IJCAI'15, AAAI Press, 2015, pp. 1376–1382.
[10] K. Filippova, E. Alfonseca, C.A. Colmenares, L. Kaiser, O. Vinyals, Sentence compression by deletion with lstms, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 360–368.
[11] S. Banerjee, P. Mitra, K. Sugiyama, Multi-document abstractive summarization using ILP based multi-sentence compression, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), 2015, pp. 1208–1214.
[12] J. Niu, H. Chen, Q. Zhao, L. Su, M. Atiquzzaman, Multi-document abstractive summarization using chunk-graph and recurrent neural network, in: 2017 IEEE International Conference on Communications (ICC), 2017, 2017, pp. 1–6.
[13] E. Linhares Pontes, S. Huet, J.-M. Torres-Moreno, A.C. Linhares, Cross-language text summarization using sentence and multi-sentence compression, in: M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, F. Meziane (Eds.), Natural Language Processing and Information Systems, Springer International Publishing, Cham, 2018, pp. 467–479.
[14] N.-T. Tran, V.-T. Luong, N.L.-T. Nguyen, M.-Q. Nghiem, Effective attention-based neural architectures for sentence compression with bidirectional long short-term memory, in: Proceedings of the Seventh Symposium on Information and Communication Technology, in: SoICT '16, ACM, New York, NY, USA, 2016, pp. 123–130.
[15] E. Linhares Pontes, S. Huet, T. Gouveia da Silva, A.C. Linhares, J.-M. Torres-Moreno, Multi-sentence compression with word vertex-labeled graphs and integer linear programming, in: Proceedings of TextGraphs-12: The Workshop on Graph-Based Methods for Natural Language Processing, Association for Computational Linguistics, 2018.
[16] A. Leuski, C.-Y. Lin, L. Zhou, U. Germann, F.J. Och, E. Hovy, Cross-lingual C*ST*RD: English access to hindi information, 2 (3) (2003) 245–269.
[17] C. Orasan, O.A. Chiorean, Evaluation of a cross-lingual romanian-english multi-document summariser, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, D. Tapias (Eds.), Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008, http://www.lrec-conf.org/proceedings/lrec2008/.
[18] X. Wan, H. Li, J. Xiao, Cross-language document summarization based on machine translation quality prediction, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, in: ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 917–926.
[19] F. Boudin, S. Huet, J. Torres-Moreno, A graph-based approach to cross-language multi-document summarization, Polibits 43 (2011) 113–118.
[20] X. Wan, Using bilingual information for cross-language document summarization, in: ACL, The Association for Computer Linguistics, 2011, pp. 1546–1555.
[21] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, C. Monz, M. Negri, M. Post, R. Rubino, L. Specia, M. Turchi, Findings of the 2017 conference on machine translation (wmt17), in: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 169–214.
[22] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.
[23] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: Association for Computational Linguistics (ACL) System Demonstrations, 2014, 2014, pp. 55–60.
[24] N. Kulkarni, M.A. Finlayson, Jmwe: A java toolkit for detecting multi-word expressions, in: Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World, in: MWE '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 122–124.
[25] Y. Zhao, X. Shen, H. Senuma, A. Aizawa, A comprehensive study: Sentence compression with linguistic knowledge-enhanced gated neural network, Data Knowl. Eng. 117 (2018) 307–318.
[26] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[27] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, CoRR, abs/1409.0473.
[28] K. Filippova, Multi-sentence compression: Finding shortest paths in word graphs, in: International Conference on Computational Linguistics (COLING), Tsinghua University Press, 2010, pp. 322–330.
[29] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[30] E. Linhares Pontes, J.-M. Torres-Moreno, S. Huet, A.C. Linhares, A new annotated portuguese/spanish corpus for the multi-sentence compression task, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.
[31] G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, V. Varma, TAC2011 multiling pilot overview, in: Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011.
[32] C.-Y. Lin, Rouge: a package for automatic evaluation of summaries, in: Proceedings of the Workshop on Text Summarization Branches Out (was 2004), 2004, pp. 74–81.
[33] F. Boudin, E. Morin, Keyphrase extraction for n-best reranking in multi-sentence compression, in: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, 2013, pp. 298–305.

[34] M. Galley, K. McKeown, Lexicalized markov grammars for sentence compression, in: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Association for Computational Linguistics, 2007, pp. 180–187.
[35] J. Clarke, M. Lapata, Modelling compression with discourse constraints, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning (EMNLP-CoNLL-2007), Prague, Czech Republic, 2007, pp. 1–11.
[36] E. Linhares Pontes, S. Huet, J.-M. Torres-Moreno, A multilingual study of compressive cross-language text summarization, in: 17th Mexican International Conference on Artificial Intelligence (MICAI), Springer, Guadalajara, Mexico, 2018.

Elvys Linhares Pontes received the Engineer's degree in Computer Science from Universidade Federal do Ceará, Sobral, Brazil, in 2013; Master's degree in Computer Science from Universidade Federal do Ceará, Sobral, Brazil, in 2015; and the Ph.D. degree in Computer Science from the Avignon Université, France in 2018. He is currently a postdoctoral researcher at the L3i, University of La Rochelle, France. His research interests include named entity linking and recognition, text summarization, multi-sentence compression, information retrieval, and machine learning for natural language processing.

Stéphane Huet received the Engineer's degree in Computer Science from INSA, Rennes, France, in 2004, and the Ph.D. degree in Computer Science from the Université de Rennes 1, France in 2007. He worked as a postdoctoral researcher at the DIRO, Université de Montréal, Canada between 2008 and 2010. He is currently an Associate Professor at the LIA, Avignon Université, France. His research interests include machine translation, dialog systems and machine learning for natural language processing.

Juan-Manuel Torres-Moreno is an associate professor at Laboratoire Informatique d'Avignon (LIA) of Université d'Avignon et des Pays de Vaucluse. He has received its Ph.D. in computer science from INP-Grenoble in 1998. He was Post-Doc researcher in the laboratory LANCI (UQAM, Montréal) and assistant professor for 3 years at Polytechnique Montréal (Québec, Canada). His research is oriented to develop algorithms in the domains of Natural Language Processing, Information Retrieval and Machine Learning. These approaches include techniques issued from statistical and symbolic paradigms.

Andréa Carneiro Linhares is a former faculty member of the computer engineering department at the Federal University of Ceará (UFC, Brazil). Linhares completed a Ph.D. in Computer Science at Avignon University (UAPV, France), a research Master and a bachelor in Computer Science, both in Brazil. In 2012, she did a post doctorate funded by UFC at the Computing Laboratory of Avignon University where she improved her knowledge in Natural Language Processing (NLP). Her experience in Computer Science is focused on Operations Research, working mainly on the following topics: combinatorial optimization, graph and algorithms, automatic summarization (NLP), frequency assignment problems and numerical methods. She has been part of international collaborations since 2010, working on multidisciplinary projects. Nowadays, she is a member of the Elites, Networks and Power in modern China project, working on the research, development and coding of new and existing algorithms, tools and technologies to design scientific applications.