Inf Retrieval (2013) 16:629–656
DOI 10.1007/s10791-012-9210-3

Cross-lingual training of summarization systems using annotated corpora in a foreign language

Marina Litvak · Mark Last

Received: 9 October 2011 / Accepted: 18 August 2012 / Published online: 6 September 2012

Abstract The increasing trend of cross-border globalization and acculturation requires text summarization techniques to work equally well for multiple languages. However, only some of the automated summarization methods can be defined as "language-independent," i.e., not based on any language-specific knowledge. Such methods can be used for multilingual summarization, defined in Mani (Automatic summarization. Natural language processing. John Benjamins Publishing Company, Amsterdam, 2001) as "processing several languages, with a summary in the same language as input", but, their performance is usually unsatisfactory due to the exclusion of language-specific knowledge. Moreover, supervised machine learning approaches need training corpora in multiple languages that are usually unavailable for rare languages, and their creation is a very expensive and labor-intensive process. In this article, we describe cross-lingual methods for training an extractive single-document text summarizer called MUSE (MUltilingual Sentence Extractor)—a supervised approach, based on the linear optimization of a rich set of sentence ranking measures using a Genetic Algorithm. We evaluated MUSE's performance on documents in three different languages: English, Hebrew, and Arabic using several training scenarios. The summarization quality was measured using ROUGE-1 and ROUGE-2 Recall metrics. The results of the extensive comparative analysis showed that the performance of MUSE was better than that of the best known multilingual approach (TextRank) in all three languages. Moreover, our experimental results suggest that using the same sentence ranking model across languages results in a reasonable summarization quality, while saving considerable annotation efforts for the end-user. On the other hand,

We evaluated several state-of-the-art summarizers—SUMMA, MEAD, Microsoft Word AutoSummary and TextRank—on the DUC 2002 corpus. Our results showed that TextRank performed best. In addition, TextRank can be considered language-independent as long as it does not perform any morphological analysis.

M. Litvak (✉)
Sami Shamoon Academic College of Engineering, 84100 Beer-Sheva, Israel
e-mail: litvakm@bgu.ac.il

M. Last
Ben Gurion University of the Negev, 84105 Beer-Sheva, Israel
e-mail: mlast@bgu.ac.il

using parallel corpora generated by machine translation tools may improve the performance of a MUSE model trained on a foreign language. Comparative evaluation of an alternative optimization technique—Multiple Linear Regression—justifies the use of a Genetic Algorithm.

Keywords Multilingual summarization · Genetic Algorithm · Cross-lingual training

1 Introduction

Document summaries should use a minimum number of words to express a document's main ideas. As such, high quality summaries can significantly reduce the constant information overload of many professionals in a variety of fields, assist in the automated classification and filtering of documents, and increase search engines precision.

Automated summarization methods can be categorized as either *statistic*-based, which use either the classic vector space model or graph representations, or as *semantic*-based, which employ ontologies and language-specific knowledge. Both categories can contain *supervised*, or *corpus-based*, machine-learning techniques, as well as *unsupervised* approaches. Automated summarization methods can use different levels of linguistic analysis: *morphological, syntactic, semantic* and *discourse/pragmatic* (Mani 2001).

Although the summary quality is expected to improve when a summarization technique includes language-specific knowledge, the dependence on such knowledge impedes the use of the same summarizer for multiple languages. On the other hand, the publication of information on the Internet in an ever-increasing variety of languages[1] dictates the importance of developing multilingual summarization approaches. Thus, there is a particular need for language-independent statistical techniques that can be readily applied to texts in any language without depending on language-specific linguistic tools.

This work is focusing on a multi-lingual summarization: we evaluate here two different approaches to the cross-lingual training of a supervised algorithm for single-document summarization—MUSE (Litvak et al. 2010b). Summarization with MUSE is considered an optimization problem, and a Genetic Algorithm (GA) is applied in order to find an optimal weighted linear combination of 31 statistical sentence features that are all language-independent. Generally speaking, our methodology described in Litvak et al. (2010b) is about finding a sentence ranking model based on a linear combination of some sentence features—by applying a Genetic Algorithm. The induced model can be applied for summarizing documents in the same or different language/genre.

We have performed our evaluation experiments on three monolingual corpora of English, Hebrew, and Arabic documents, and six parallel corpora resulted from the pairwise machine translation of each corpus from its original language to the other two. The experiments were aimed at evaluating our approach in both mono-lingual and multilingual environments as well as comparing it to the state-of-the-art summarization methods and optimization approaches.

This article is organized as follows. The next section describes the related work in extractive summarization. Section 3 describes MUSE, the GA-based approach to multilingual single-document extractive summarization, and possible scenarios for a cross-lingual training. Section 4 presents our experimental results for the multilingual

[1] Authors of (Gulli and Signorini 2005) used Web searches in 75 different languages to estimate the size of the Web as of the end of January 2005.

summarization task and cross-lingual training of MUSE. Our conclusions and suggestions for future work comprise the final section.

2 Related work

Extractive summarization is aimed at the selection of a subset of the most relevant fragments from a source text into the summary. The fragments can be paragraphs (Salton et al. 1997), sentences (Luhn 1958), keyphrases (Turney 2000; Litvak et al. 2011) or keywords (Litvak and Last 2008). Extractive summarization usually consists of *ranking*, where each fragment of a summarized text gets a relevance score, and *extraction*, where the top-ranked fragments are gathered into a summary, according to their appearance in the original text. Statistical methods for calculating the relevance score of each fragment can be categorized into several categories: *cue*-based (Edmundson 1969), *keyword*- or *frequency*-based (Luhn 1958; Edmundson 1969; Neto et al. 2000; Steinberger and Jezek 2004; Kallel et al. 2004; Vanderwende et al. 2007), *title*-based (Edmundson 1969; Teufel and Moens 1997), *position*-based (Baxendale 1958; Edmundson 1969; Lin and Hovy 1997; Nobata et al. 2001) and *length*-based (Nobata et al. 2001). In our approach, we use 31 language-independent sentence features from various categories.

Considered the first work on sentence scoring for automated text summarization, the seminal paper by Luhn (1958) based the significance factor of a sentence on the frequency and the relative positions of significant words within a sentence. Edmundson (1969) tested different linear combinations of four scoring features for ranking sentences—*cue*, *key*, *title* and *position*—to identify the one with the best performance for a training corpus. Linear combinations of several statistical sentence ranking features were also applied in the MEAD (Radev et al. 2001) and SUMMA (Saggion et al. 2003) approaches, both of which use the vector space model for text representation and a set of predefined or user-specified weights for a combination of *position*, *frequency*, *title*, and *centroid*-based (MEAD) features. Goldstein et al. (1999) integrated linguistic and statistical features. In none of these works, however, did the researchers attempt to find the optimal weights for the best linear combination. Later, attempts to find the best combination had been done using machine learning techniques, like in Wong et al. (2008), where supervised and semi-supervised learning approaches were applied to various sentence features. Different groups of features from four categories were manually constructed and evaluated, and, finally, 14 features from three categories were found as the best combination. In our work, we continue these attempts by the supervised learning of the best linear combination of 31 sentence features from a training set of annotated documents. Our approach applies a global search technique to a full set of features and does not require to construct different combinations manually.

Some authors reduced the summarization process to an optimization or a search problem. Hassel and Sjobergh (2006) used a standard hill-climbing algorithm to build summaries that maximize the score for the total impact of the summary. A summary consisting of the first sentences from the document was used as a starting point for the search, and all neighbors (summaries that can be created by simply removing one sentence and adding another) were examined, looking for a better summary. Aker et al. (2010) used the $A*$ search algorithm to find the best extractive summary up to a given length, which is both optimal and computationally efficient. Ouyang et al. (2011) applied regression models to query-focused multi-document summarization, where they used Support Vector Regression (SVR) to estimate the importance of a sentence in a document set to be summarized through a set of pre-defined features. In our work, we use a Genetic Algorithm

(GA), which is known as a prominent search and optimization method (Goldberg 1989), for optimizing a linear combination of multiple sentence features.

Alfonseca and Rodriguez (2003), Kallel et al. (2004) and Liu et al. (2006b) used GAs in order to find sets of sentences that maximize summary quality metrics, starting from a random selection of sentences as the initial population. In this setting, however, the high computational complexity of GAs is a disadvantage. To choose the best summary, multiple candidates should be generated and evaluated for each document (or document cluster). Following a different approach, Turney (2000) used a GA to learn an optimized set of parameters for a keyword extractor embedded in the Extractor tool.[2] Orăsan et al. (2000) enhanced the preference-based anaphora resolution algorithms by using a GA to find an optimal set of values for the outcomes of 14 indicators and apply the optimal combination of values obtained from data on one text to a different text. With such an approach, training may be the only time-consuming phase in the process. The detailed description of our approach to using a GA for optimizing the sentence feature weights can be found in the next section.

All corpus-based approaches have one common problem—they need to be retrained for each new language and genre. However, preparing annotated corpora for multiple languages is a very labor-intensive and time-consuming process, especially for rare languages. Nowadays, the use of parallel corpora is very popular in different areas of information retrieval and computational linguistics, including cross-lingual summarization. The researchers in Wan et al. (2010) have adopted the Late Translation (LateTrans) strategy, using machine translation, for cross-lingual summarization. They evaluated the translation quality of each sentence in the English-to-Chinese summarization of a given document or a document set, and, finally, the English sentences with high translation quality and high informativeness were selected and translated to form the Chinese summary. In this article we show empirically that using parallel corpora can be helpful for training corpus-based summarization techniques when no training data for a new language is available.

Various text representation models for summarized documents had been utilized across different approaches. Today, graphs are becoming increasingly popular, due to their ability to enrich the document model with syntactic and semantic relations. Erkan and Radev (2004) and Mihalcea (2005) introduced LexRank and TextRank, respectively—algorithms for unsupervised extractive summarization that rely on the application of iterative graph-based ranking algorithms, such as PageRank (Brin and Page 1998) and HITS (Kleinberg 1999). Their methods represent a document as a graph of sentences interconnected by similarity relations. Various similarity functions can be applied: cosine similarity as in LexRank (Erkan and Radev 2004), simple overlap as in TextRank (Mihalcea 2005), or other functions. Edges representing the similarity relations can be weighted (Mihalcea 2005) or unweighted (Erkan and Radev 2004): two sentences are connected if their similarity is above some predefined threshold value. Wan (2008) applied the graph-based ranking algorithm based on each kind of sentence relationship for generic multi-document summarization, and integrated the relevance of the sentences to the specified topic into the graph-ranking based method for topic-focused multi-document summarization. In MUSE, we use two graph-based models, which are based on sentence and word segmentation, respectively.

It is worth noting that our work is aimed at a *generic* summarization representing the author's point of view that is different from a *query-based* summarization focusing on material of interest to the user (Hovy 2001). While in generic summarization the only input

[2] http://www.extractor.com/.

for a system is a document (or documents) to summarize, in query-based summarization a query expressing the user's interest has to be provided. A query-based summary must contain the information relevant to a given query.

3 MUSE: MUltilingual Sentence Extractor

MUltilingual Sentence Extractor is a supervised learning approach to language-independent extractive summarization, where the best set of weights for a linear combination of sentence scoring methods is found by a genetic algorithm trained on a collection of document summaries (see Algorithm 1). Formally, the model for sentence scoring can be expressed by the following formula:

$$Score = \sum w_i \times r_i$$

where r_i is the value of *i*th sentence feature (one of 31 described below) and w_i is its weight in the linear combination.

The weighting vector thus obtained is to be used for sentence scoring in future summarizations. The sentences with the highest score are then selected for the summary, according to the greedy approach presented in Algorithm 2.

Since most sentence scoring methods have a linear computational complexity, only the training phase of our approach is time-consuming.

Figure 1 depicts the flowchart of the proposed approach. It consists of two main modules: the *training module* activated offline, and the *summarization module* operating online. Both modules utilize three different representations of documents: one vector-based

Algorithm 1 Step 1: Training

```
Input: Gold Standard - a corpus of summarized documents $D$, $N$ sentence features
Output: A weighted model $W$ - vector of weights for each of $N$ features
  Step 1.1: Compute $M$ - sentence-score matrix
  for all $d \in D$ do
    for all sentences $s \in d$ do
      Calculate $N$ features
      Add a row of feature scores for $s$ into $M$
    end for
  end for
  Step 1.2: Compute a vector $W$ of features weights
   Run a Genetic Algorithm on $M$, given $D$:
   Initialize a population $P$
  repeat
    for all solutions $g \in P$ do
      Generate a summary $a$
      Evaluate $a$ by ROUGE on summaries of $D$
    end for
     Select the best solutions $G$
     Generate a new population $P$ from $G$
  until convergence - no better solutions are found
  return a vector $W$ of weights - output of GA
```

Algorithm 2 Step 2: Summarizing a new document

```
Input: A document $d$, maximal summary length $L$, a trained weighted model $W$
Output: A set of $n$ sentences, which were top-ranked by the algorithm
  Step 2.1: Compute a score of each sentence
  for all sentences $s \in d$ do
      Calculate $N$ features
      Calculate a score as a linear combination according to $W$
  end for
  Step 2.2: Compile the document summary
   Let $E = \emptyset$ be a summary of $d$
  repeat
      Get the top ranked sentence $s_i$
      $E = E \cup s_i$
  until $E$ exceeds max length $L$
  return $E$
```

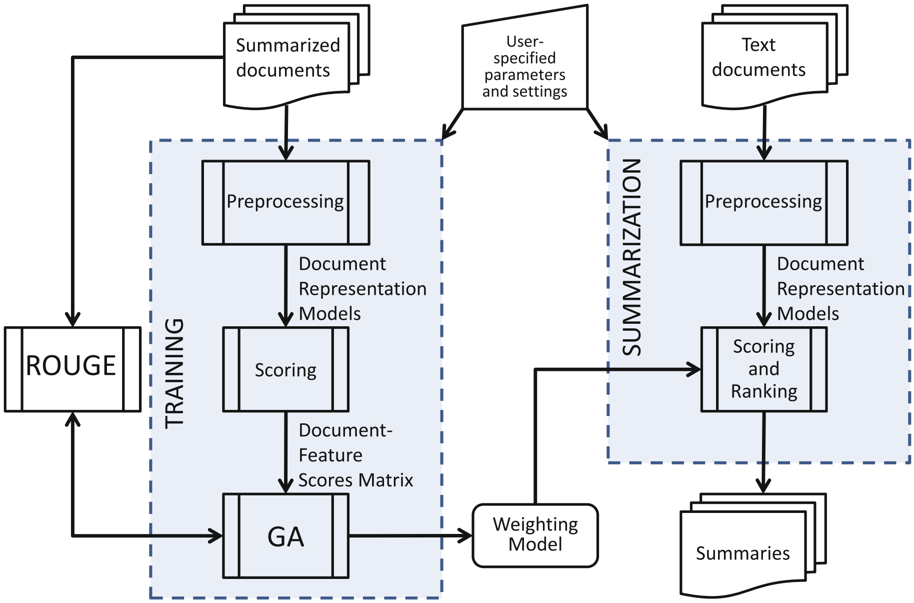


Fig. 1 MUSE summarization flowchart

and two graph-based (see Sect. 3.2). The *preprocessing sub-module* is responsible for constructing each representation, and it is embedded in both modules. Algorithms 1 and 2 contain the pseudo-code for two independent phases of MUSE: training and summarization, respectively.

The *training module* receives as input a corpus of documents, each accompanied by one or several gold-standard summaries—abstracts or extracts[3]—compiled by human

[3] Both abstracts and extracts are brief synopses of a document, while extracts are composed of exact portions of text extracted from the source document, and abstracts involve paraphrasing sections of the source document.

assessors. The set of documents may be either monolingual or multilingual and their summaries have to be in the same language as the original text. As a second parameter, the module obtains a user-specified set of sentence features computed by the system. Then, the training module applies a genetic algorithm to sentence-feature matrices of precomputed sentence scores for each input feature with the purpose of finding the best linear combination of features using ROUGE[4] as a fitness function. The output of the training module is a vector of weights for user-specified sentence ranking features.

The *summarization module* performs an on-line summarization of input text/texts. Each sentence of an input text document obtains a relevance score according to the trained model, and the top ranked sentences are extracted to the summary in their original order. To avoid duplicate content, a new sentence is added if and only if it is not similar to the previously selected sentences. In our experiments, we used cosine similarity measure with a threshold of 0.8. The length of resulting summaries is limited by a user-specified value (maximum number of words or sentences in the text extract or a maximum extract-to-text ratio). The summarization module is expected to use the model trained on the same language as input texts. However, if such model is not available (no annotated corpus in the text language), the user can choose from the following: (1) a model trained on some other language/corpus can be used (in Sect. 4 we explore whether the same model can be efficiently used across different languages), or (2) a model can be trained on a parallel corpus generated by a machine translation tool.

The *preprocessing submodule* performs the following tasks: (1) sentence segmentation, (2) word segmentation, (3) stopwords removal,[5] (4) vector space model construction using *tf* and/or *tf-idf* weights, (5) a word-based graph representation construction, (6) a sentence-based graph representation construction, and (7) document metadata extraction. The outputs of this submodule are: sentence segmented text, vector space model, and the document graphs. Both modules—training and summarization—use all three representation modules for calculation of sentence features. It is worthwhile to note that, proper sentence segmentation is crucial for the quality of extractive summarization results. Since sentence and word segmentation are language-dependent,[6] these parts should be integrated and configured for each language by the end-user of our system. So far, we have used the sentence splitter provided with the MEAD summarizer (Radev et al. 2001) for English sentences,[7] and a simple splitter that can split the text at periods, exclamation points, or question marks for the Hebrew and Arabic texts. In the future we intend to utilize a fully language-independent technique for text segmentation based on n-grams.

3.1 Language-independent sentence scoring features

MUltilingual Sentence Extractor is aimed at identifying the best linear combination of language-independent sentence scoring features. Table 1 shows the complete list of 31

[4] In this article we report using ROUGE-1 and ROUGE-2 Recall metrics. We trained and tested our system against the same metric in each set of experiments.

[5] This stage is optional, according to user's setting. In our experiments, we used stopword removal with the English and Hebrew corpora (unlike Arabic).

[6] In languages like Hebrew and Arabic, the period marks indicate the end of the sentence, which is not always true for English texts. In German and Chinese, spaces do not necessarily separate words like in most other languages.

[7] Although the same set of splitting rules may be used for many different languages, separate splitters were used because the MEAD splitter tool is restricted to European languages.

Table 1 Sentence scoring features (Litvak et al. 2010b)

Name	Description	Source
POS_F	Closeness to the beginning of the document: $\frac{1}{i}$	Edmundson (1969)
POS_L	Closeness to the end of the document: i	Baxendale (1958)
POS_B	Closeness to the borders of the document: $max(\frac{1}{i}, \frac{1}{n-i+1})$	Lin and Hovy (1997)
LEN_W	Number of *words* in the sentence	Nobata et al. (2001)
LEN_CH	Number of *characters* in the sentence	
LUHN	$max_{i \in \{clusters(S)\}}\{CS_i\}$, $CS_i = \frac{W_i^2}{N_i}$	Luhn (1958)
KEY	Sum of the keywords frequencies: $\sum_{t \in \{Keywords(S)\}} tf(t)$	Edmundson (1969)
COV	Ratio of keywords number (Coverage): $\frac{\lvert Keywords(S) \rvert}{\lvert Keywords(D) \rvert}$	Liu et al. (2006a)
TF	Average term frequency for all sentence words: $\frac{\sum_{t \in S} tf(t)}{N}$	Vanderwende et al. (2007)
TFISF	$\sum_{t \in S} tf(t) \times isf(t)$, $isf(t) = 1 - \frac{log(n(t))}{log(n)}$, $n(t)$ is the number of sentences containing t	Neto et al. (2000)
SVD	Length of a sentence vector in $\Sigma^2 \cdot V^T$ after computing Singular Value Decomposition of a term by sentences matrix $A = U\Sigma V^T$	Steinberger and Jezek (2004)
TITLE_O	Overlap similarity to the title: $sim(S, T) = \frac{\lvert S \cap T \rvert}{min\{\lvert S \rvert, \lvert T \rvert\}}$	Edmundson (1969)
TITLE_J	Jaccard similarity to the title: $sim(S, T) = \frac{\lvert S \cap T \rvert}{\lvert S \cup T \rvert}$	
TITLE_C	Cosine similarity to the title: $sim(S, T) = cos(\mathbf{S}, \mathbf{T}) = \frac{\mathbf{S} \bullet \mathbf{T}}{\lvert \mathbf{S} \rvert \lvert \mathbf{T} \rvert}$	
D_COV_O	Overlap similarity to the document complement $sim(S, D - S) = \frac{\lvert S \cap (D - S) \rvert}{min\{\lvert S \rvert, \lvert D - S \rvert\}}$	Litvak et al. (2010b)
D_COV_J	Jaccard similarity to the document complement $sim(S, D - S) = \frac{\lvert S \cap (D - S) \rvert}{\lvert S \cup D - S \rvert}$	
D_COV_C	Cosine similarity to the document complement $cos(\mathbf{S}, \mathbf{D} - \mathbf{S}) = \frac{\mathbf{S} \bullet (\mathbf{D} - \mathbf{S})}{\lvert \mathbf{S} \rvert \lvert \mathbf{D} - \mathbf{S} \rvert}$	
LUHN_DEG **KEY_DEG** **COV_DEG**	Graph-based extensions of LUHN, KEY and COV measures respectively. Node degree is used instead of a word frequency: words are considered significant if they are represented by nodes having a degree higher than a predefined threshold	
DEG	Average degree for all sentence nodes: $\frac{\sum_{i \in \{words(S)\}} Deg_i}{N}$	
GRASE	Frequent sentences from *bushy* paths are selected. Each sentence in the *bushy* path gets a domination score that is the number of edges with its label in the path normalized by the sentence length. The relevance score for a sentence is calculated as the sum of its domination scores over all paths.	
LUHN_PR **KEY_PR** **COV_PR**	Graph-based extensions of LUHN, KEY and COV measures respectively. Node PageRank score is used instead of a word frequency: words are considered significant if they are represented by nodes having a PageRank score higher than a predefined threshold	
PR	Average PageRank for all sentence nodes: $\frac{\sum_{t \in S} PR(t)}{N}$	

Table 1 continued

Name	Description	Source
TITLE_E_O	Overlap-based edge matching between title and sentence graphs	
TITLE_E_J	Jaccard-based edge matching between title and sentence graphs	
D_COV_E_O	Overlap-based edge matching between sentence and a document complement graphs	
D_COV_E_J	Jaccard-based edge matching between sentence and a document complement graphs	
ML_TR	Multilingual version of TextRank without morphological analysis: sentence score equals PageRank (Brin and Page 1998) rank of its node: $WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$	Mihalcea (2005)

sentence features used in this article. Each feature description includes a reference to the original work where the method was proposed for extractive summarization. Several methods were proposed by us in Litvak et al. (2010b). Formulas incorporate the following notation: a sentence is denoted by S, a text document by D, the total number of words in S by N, the total number of sentences in D by n, the sequential number of S in D by i, and the in-document term frequency of the term t by $tf(t)$. In the *LUHN* method, W_i and N_i are the number of keywords and the total number of words in the ith cluster, respectively, whereas clusters are sentence portions bracketed by keywords, i.e., frequent, non-common words.[8]

Due to the multilingual focus of our work, *exact* word matching was used in all similarity-based methods. From the same reason, we kept two kinds of length features: number of words and number of characters[9] in the sentence.

Figure 2 demonstrates the taxonomy of the 31 features listed in Table 1. The features are divided into three main categories—*structure*-, *vector*-, and *graph*-based—according to the type of text representation they use, where each category is divided into sub-categories according to the main calculating criteria. For example, the "graph-based" category contains all features that use graph representation module (word- and sentence-based), and its "pagerank" sub-category combines features based on the eigenvector node (standing for word or sentence) centrality. Features that require pre-defined threshold values are marked with a cross and listed in Table 2 together with the average threshold values obtained after method evaluation on English, Hebrew, and Arabic corpora. Each feature was evaluated on three corpora, with different thresholds $t \in [0, 1]$ (only values with one decimal digit were considered). The threshold values that resulted in the best ROUGE scores across three corpora, as a result of training on the entire corpus,[10] were selected. A threshold of 1 means that all terms are considered, while a value of 0 means that only terms with the highest absolute score of *tf*, *degree*, or *pagerank* (depends on a feature) are considered.

Section 3.3 describes our application of a Genetic Algorithm to the summarization task.

[8] Luhn's experiments suggest an optimal limit of 4 or 5 non-significant words between keywords.

[9] This variation is more appropriate for multilingual processing due to different rules of tokenization in different languages—for example, English versus German.

[10] Without cross-validation.

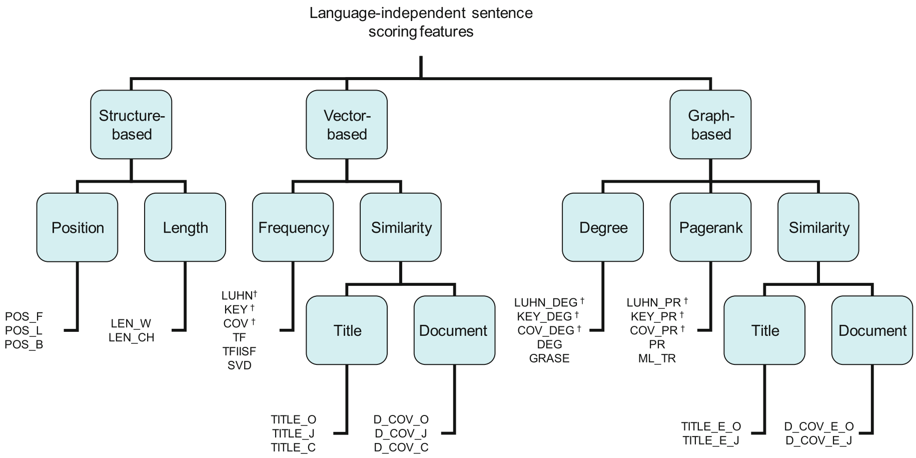


Fig. 2 Taxonomy of language-independent sentence scoring features (Litvak et al. 2010b)

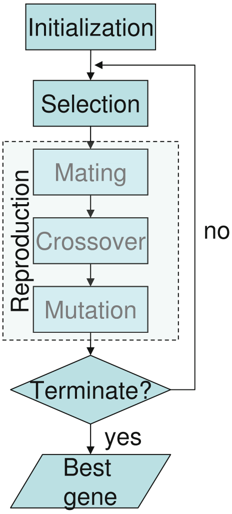
Table 2 Selected thresholds for threshold-based scoring methods (Litvak et al. 2010a)

Method	Threshold
LUHN	0.9
LUHN_DEG	0.9
LUHN_PR	0.0
KEY	[0.8, 1.0]
KEY_DEG	[0.8, 1.0]
KEY_PR	[0.1, 1.0]
COV	0.9
COV_DEG	[0.7, 0.9]
COV_PR	0.1

3.2 Text representation models

The vector-based scoring methods listed in Table 1 use *tf* or *tf-idf* term weights to evaluate sentence importance. In contrast, representation used by the graph-based methods (all except TextRank) is based on the word-based graph representation models described in Schenker et al. (2004). Schenker et al. (2005) showed that such graph representations can outperform the vector space model on several text mining tasks. In the word-based graph representation used in our work, nodes represent unique terms (distinct words) and edges represent order-relationships between two terms. There is a directed edge from *A* to *B* if an *A* occurrence immediately precedes the *B* occurrence in any sentence of the document. Contrary to Schenker et al. (2005), we have labeled each edge with the IDs of sentences that contain both words in the specified order. For the TextRank score calculation (denoted by ML_TR in Table 1), we build a sentence-based graph representation where nodes stand for sentences and edges for similarity relationships.

Fig. 3 GA flowchart



3.3 Optimization: learning the best linear combination

We found the best linear combination of the features listed in Table 1 using a Genetic Algorithm (GA). GAs are categorized as global search heuristics. Figure 3 (Litvak et al. 2010b) shows a simplified flowchart of a Genetic Algorithm.

A typical genetic algorithm requires (1) a genetic representation of the solution domain, (2) a fitness function to evaluate the solution domain, and (3) selection and reproduction rules.

We represent each solution as a vector of weights for a linear combination of sentence scoring features—real-valued numbers in an unlimited range, normalized in such a way that they sum up to 1. The vector size is fixed and it equals to the number of features used in the combination.

Defined over the genetic representation, the fitness function measures the quality of the represented solution. We use ROUGE-1 and ROUGE-2, Recall (Lin and Hovy 2003) as fitness functions for measuring summarization quality—similarity with gold standard summaries, which should be *maximized* during the training (optimizing procedure). For a training set, we use an annotated corpus of summarized documents, where each document is accompanied by several human-generated summaries—abstracts or extracts.[11]

Below each phase of the optimization procedure is described in detail.

Initialization: GA explores only a small part of the search space if the population is too small, whereas it slows down if there are too many solutions. We start from $N = 500$ randomly generated genes/solutions as an initial population, that empirically was proven as a good choice during our experiments. Each gene is represented by a weighting vector $v_i = w_1, \ldots, w_D$ with a fixed number D of elements that equals to the number of sentence features[12] used in linear combination. All elements are generated from a standard normal

[11] The average number and type of gold standard summaries are different in different corpora. The details are reported in Sect. 4.

[12] In our experiments we used $D = 31$, that is the number of sentence features used for finding the best liner combination.

distribution, with $\mu = 0$ and $\sigma^2 = 1$, and normalized to sum up to 1. For this solution's representation, a negative weight, if it occurs, can be considered as a "penalty" for the associated feature.

Selection: During each successive generation, a proportion of the existing population is selected to breed a new generation. We use a truncation selection method that rates the fitness of each solution and selects the best fifth (100 out of 500) of the individual solutions, i.e., getting the maximal ROUGE value. In such manner, we discard "bad" solutions and prevent them from reproducing. In addition, we use *elitism*—a method that prevents losing the best found solution in the population by copying it to the next generation.

Reproduction: At this stage, new genes/solutions are introduced into the population, i.e., new points in the search space are explored. These new solutions are generated from those selected through the following genetic operators: *mating, crossover*, and *mutation*.

In *mating*, a pair of "parent" solutions is randomly selected, and a new solution (child) is created using *crossover* and *mutation*, which are the most important parts of a genetic algorithm. The GA performance is influenced mainly by these two operators. New parents are selected for each new child, and the process continues until a new population of solutions of appropriate size N is generated.

Crossover is performed under the assumption that new solutions can be improved by re-using the good parts of old solutions. However it is beneficial to keep and transfer some part of population from one generation to the next. Our *crossover* operator includes a probability (80 %) that a new and different offspring solution is generated by calculating the weighted average of two "parent" vectors according to Vignaux and Michalewicz (1991). Formally, a new vector v is created from two vectors v_1 and v_2 according to the formula $v = \lambda * v_1 + (1 - \lambda) * v_2$ (we set $\lambda = 0.5$). There is a probability of 20 % that the offspring is a duplicate of one of its parents. The reason for allowing dupicates in some cases is a balancing between exploration and exploitation—a very high crossover rate relative to selective pressure, given high initial variability and a very large population, may turn evolution into a random search (Goldberg 1989).

Mutation in GAs functions both in preserving the existing diversity and introducing new variation. It is aimed at preventing the GA from falling into a local extremum, but it should not be applied too often, due to the danger of transforming the GA into a random search. The mutation operator introduced here includes a probability (3 %) that an arbitrary weight in a vector would be changed by a uniformly randomized factor in the range of $[-0.3, 0.3]$ around its original value.

Termination: The generational process is repeated until a termination condition—a plateau of solution/combination fitness such that successive iterations no longer produce significantly better results—has been reached. In our implementation, just one iteration must show no significant improvement in the best individual fitness before the termination. The minimal improvement in our experiments was set to $\epsilon = 1.0E - 21$.[13]

3.4 Training scenarios

The training of MUSE can be performed according to *monolingual* and/or *cross-lingual* scenarios, depending on either of the following:

[13] Since we measure our fitness as a ROUGE score, theoretically, it may vary from 0 to 1. Practically, it varied from 0.2 to 0.449 (the best fitness in training) for ROUGE-1 on English corpus.

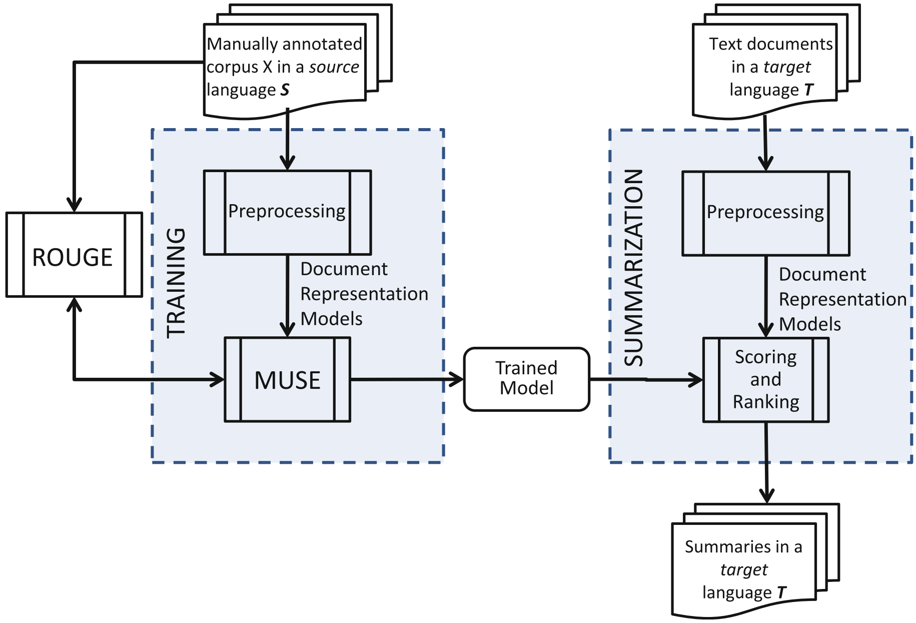

Manually annotated corpus X in a *source* language ***S***
ROUGE
TRAINING
Preprocessing
Document Representation Models
MUSE
Trained Model
Text documents in a *target* language ***T***
SUMMARIZATION
Preprocessing
Document Representation Models
Scoring and Ranking
Summaries in a *target* language ***T***

Fig. 4 Cross-lingual training using source-language corpora

1. *A training corpus in the target language is available*. Since MUSE is language-independent, it can be trained on a corpus of summarized documents in any target language. This approach is called "monolingual training".
2. *A training corpus in the target language is not available, but there is a training corpus in a different ("source") language*. Here several options can be considered:

(a) One may train MUSE on the existing corpus and use the same trained model across different languages. Figure 4 depicts the flowchart of such an approach which is called "cross-lingual training". This approach is quite problematic since, despite the language-independency of MUSE, different languages may have different trained models. The next scenario is aimed to solve the tradeoff between expensive manual annotation and multilingual summarization performance.

(b) In order to obtain *language-oriented* trained models[14] in the case of a lack of data in the target language, one may translate a corpus from source to target language using machine translation tools, and use the parallel corpora for training. We propose a methodology for cross-lingual training of a summarization system that is based on the early translation strategy, where each document in the training corpus is translated to the target language prior to model learning. The flowchart of this scenario is depicted in Fig. 5.

[14] We assume that some of the 31 sentence features may have a different impact on the ranking model in different languages. For example, all frequency-based features may affect models in Hebrew and Arabic, where stopwords such as "on", "in", "to", etc. cannot be easily identified (unlike English), since they are prefixes of non-stop words. Stopword removal after machine translation in cross-lingual learning may help to create a language-oriented model with appropriate weights for the frequency-based features. Of course, the differences in translation quality of different language pairs may affect the summarization quality as well.

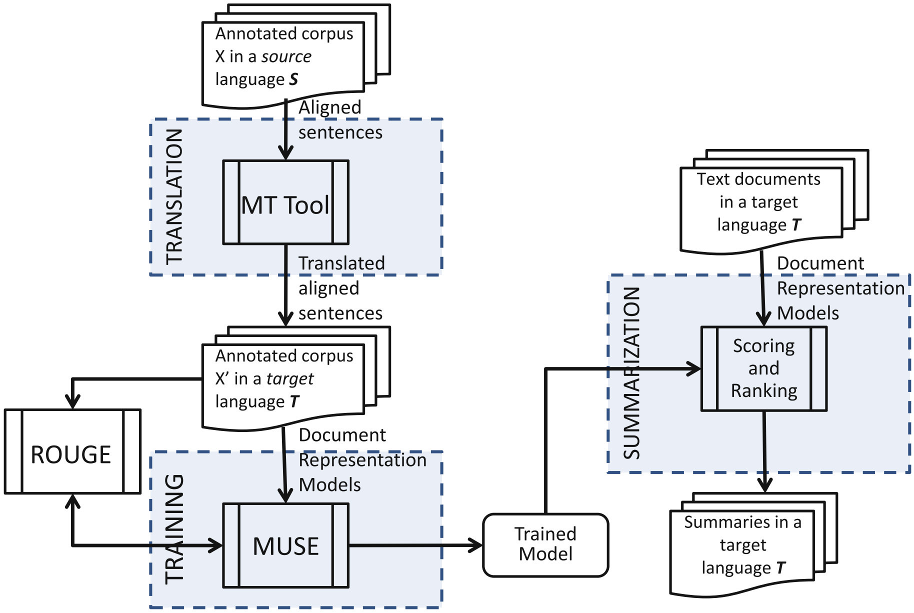

Annotated corpus X in a *source* language ***S***
Aligned sentences
TRANSLATION
MT Tool
Translated aligned sentences
Annotated corpus X' in a *target* language ***T***
ROUGE
Document Representation Models
TRAINING
MUSE
Trained Model
Text documents in a target language ***T***
Document Representation Models
SUMMARIZATION
Scoring and Ranking
Summaries in a target language ***T***

Fig. 5 Cross-lingual training using parallel corpora

Note that MUSE and ROUGE can be replaced in the figures above with *any* corpus-based summarizer and evaluation tool, respectively. Both scenarios for cross-lingual learning are generally applicable to any language-independent summarizer or language-dependent summarizer adapted to the source/target language.

3.5 Complexity issues

Assuming efficient implementation, most sentence ranking methods used by MUSE have a linear computational complexity relative to the total number of words in a document—$O(n)$. As a result, MUSE document summarization time, given a trained model, is also linear in the number of features in a combination. The training time is proportional to the number of GA iterations multiplied by the number of individuals in a population, times the fitness evaluation (ROUGE) time. On average, in our experiments the GA performed only 5–6 iterations of selection and reproduction before reaching convergence.

4 Experiments

4.1 Overview

The MUSE summarization approach and the quality of its cross-lingual training were evaluated using comparative experiments on three monolingual corpora of English, Hebrew, and Arabic texts. These languages were intentionally chosen, since they belong to distinct language families (Indo-European and Semitic languages, respectively), to ensure that the results of our evaluation would be widely generalizable. The specific goals of the experiment were:

1. Evaluate the optimal sentence scoring models induced from the corpora of summarized documents in three different languages,
2. Determine whether the same sentence scoring model could be efficiently used for extractive summarization across three different languages,
3. Determine whether using parallel corpora in cross-lingual training improves the multilingual performance of MUSE,
4. Compare the performance of the GA-based summarization method to the state-of-the-art approaches, and
5. Compare the GA performance to alternative optimization techniques, viz. Multiple Linear Regression (MLR).

The following subsections describe: our experimental setup (data, evaluation metrics and scenarios), experimental results, and their discussion.

4.2 Experimental setup

4.2.1 Corpora

The English text material used in the experiments comprised the corpus of summarized documents available for the summarization task at the Document Understanding Conference 2002 (DUC 2002). This benchmark dataset contains 533 news articles, each accompanied by two to three human-generated *abstracts* of approximately 100 words each.

For the Arabic language, we generated (in collaboration with several experts in Arabic) a new corpus compiled from 90 news articles. Each article was summarized by three native Arabic speakers selecting the most important sentences into an *extractive* summary of approximately 100 words each. All assessors were provided with the *Tool Assisting Human Assessors* (TAHA) software tool[15] that enables sentences to be easily selected and stored for later inclusion in the document extract. The agreement between assessors measured by ROUGE-1 (Lin and Hovy 2003) score shows that their summaries overlap by 75 % on average.

For the Hebrew language, we used the corpus generated as part of our experiment[16] where 120 news articles of 250–830 words each from the websites of the *Haaretz* newspaper,[17] *The Marker* newspaper,[18] and manually translated articles from *WikiNews*[19] were summarized by human assessors using the TAHA software. In total, 126 undergraduate students from the Department of Information Systems Engineering, Ben Gurion University of the Negev participated in the experiment. Each participant was randomly assigned ten different documents and instructed to choose the most important sentences in each document subject to the following constraints: (1) spend at least five minutes on each document, (2) ignore dialogs and quotations, (3) read the whole document before beginning sentence extraction, (4) ignore redundant, repetitive, and overly detailed information, and (5) remain within the minimal and maximal summary length limits (95 and 100 words, respectively). Summaries were assessed for quality by comparing each student's extract to

[15] TAHA can be provided upon request.

[16] The Hebrew corpus used in this article is an extension of the original Hebrew corpus introduced in Litvak et al. (2010b).

[17] http://www.haaretz.co.il.

[18] http://www.themarker.com.

[19] http://en.wikinews.org/wiki.

those of all the other students using the ROUGE evaluation toolkit and the ROUGE-1 metric. We discarded the summaries produced by assessors who received an average ROUGE score below 0.5, i. e. agreed with the rest of assessors in less than 50 % of cases. Also, the time spent by an assessor on each document was checked (with respect to the requirements). The final corpus of summarized Hebrew texts was compiled from the summaries of about 60 % of the assessors, with an average of five *extracts* per single document. The average ROUGE scores of the selected assessors is 54 %. The dataset is available at http://www.cs.bgu.ac.il/~litvakm/research/.

Three corpora have different characteristics of the gold standard summaries with respect to the following parameters:

- **type** of a summary: the English corpus contains *abstracts*, whereas the Hebrew and Arabic corpora both contain *extracts* (extracted sentences);
- **number** of summaries per document: the English corpus contains from two to three summaries, the Arabic has exactly three extracts, and the Hebrew corpus consists of five extracts per document on average;
- **diversity** of summaries: while the Hebrew corpus contains the most diverse summaries (each assessor summarized only ten documents from 120), the Arabic corpus has the most consistent summaries since the same three assessors summarized all corpus documents;
- **coverage** of summaries: in the Arabic and Hebrew corpora many extracts are compiled of initial sentence/sentences, while English abstracts contain information representing all sentences in the source document.

The documents from all corpora have a title as the first sentence. The parallel corpora were obtained by machine translating each one of the three monolingual corpora from the source language to two target languages using Google Translate API[20] and sentence segmentation.

4.2.2 *Evaluation metrics*

We evaluated English, Hebrew, and Arabic summaries using ROUGE-1 and ROUGE-2 metrics, described in Lin (2004b). Similar to Lin's conclusion in Lin (2004b), our results for the different ROUGE metrics were not statistically distinguishable. However, ROUGE-1 showed the largest variation across the methods and, according to the conclusion made in Lin (2004a), ROUGE-2 is a good choice in single-document summarization tasks. In the following comparisons, all results are presented in terms of the ROUGE-1 and ROUGE-2 Recall metrics. In order to use the ROUGE toolkit on Hebrew and Arabic, it was adapted to these languages by specifying the regular expressions for a single "word" using Hebrew and Arabic characters.

4.2.3 *Evaluation scenarios*

According to the goals of our experiment listed in Sect. 4.1 above, we performed the following evaluations:

1. We evaluated the *monolingual* training of MUSE on each monolingual corpus using 10-fold cross validation.

[20] The translations were obtained in October-November 2010 and May 2012.

2. We compared the MUSE approach with the following *unsupervised* summarization methods:

   (a) a multilingual version of *TextRank* (Mihalcea 2005) as the best known multilingual summarizer[21] (denoted as ML_TR in Table 1),
   (b) degree-based *Coverage* (denoted as COV_DEG in Table 1) as the best single scoring method in English corpus, and
   (c) the *Baseline* approach compiling the summaries from the initial sentences (denoted as POS_F in Table 1). The baseline approach was found to be the best single feature in Arabic and Hebrew corpora.[22]

3. As a part of the monolingual experiment, we compared the performance of two different optimization techniques used to calculate the optimal linear combination of sentence features. Since a Genetic Algorithm is known as a time- and space-consuming technique, it was compared to a common and simple optimization method—Multiple Linear Regression. For the experiment, the corpus of English summarized documents (DUC 2002) was utilized. We have calculated 31 features (see Table 1) as well as a ROUGE score for each sentence of the corpus documents,[23] where the ROUGE value represented the sentence relevance score for inclusion in the document summary. Then, the Least Squares Algorithm[24] was run, in order to estimate a multiple linear regression model—a linear combination of 31 features—with a ROUGE score as the dependent variable, predicting the sentence relevance score in the future summarization, and 31 independent predictor variables representing 31 sentence features.
4. We evaluated the quality of *cross-lingual* training with MUSE by applying the model trained on a corpus in one (source) language to documents in another (target) language.
5. The last phase of our experiment was to determine whether using parallel corpora in cross-lingual training improves the multilingual performance of MUSE. All available data (translated and target corpora, respectively) was used for training and testing the summarizer using parallel corpora. The 10-fold cross validation was applied.

Three research hypotheses were tested performing three different statistical tests formulated in Table 3. The research ("alternative") hypotheses are shown in the right column of Table 3. In order to perform the testing, the results were analyzed and compared to the null hypotheses (shown in the left column of Table 3) using paired *t* test or Wilcoxon matched-pairs signed-ranks test, according to whether the data passed the normality test using the method of Kolmogorov and Smirnov.

[21] Since the TextRank code is unavailable (we asked the authors of (Mihalcea 2005) for the TextRank code in the past), we implemented our own version according to the description in the TextRank paper.

[22] In Litvak et al. (2010b) MUSE was also compared with Microsoft Word 2007 AutoSummary Tool in English and Hebrew as a widely spread commercial summarizer. We did not apply this tool to Arabic documents, since it does not support the Arabic language.

[23] Since the DUC 2002 corpus is comprised of *abstracts*, their inclusion into human summaries could not be simply checked. Instead, the ROUGE score of each sentence was measured, simulating its relevance score for the summary.

[24] Weka software was used, http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html.

Table 3 Performed tests: alternative and null hypotheses

Test	Null hypothesis (H_0)	Alternative hypothesis (H_1)
1	MUSE does not outperform other approaches.	MUSE outperforms othe approaches.
2	Training MUSE on source-language corpora does not decrease the summarization quality.	Training MUSE on source-language corpora decreases the summarization quality.
3	Retraining MUSE on parallel corpora does not improve its performance in cross-lingual learning versus training on source-language corpora.	Using parallel corpora improves performance of cross-lingual learning versus training on source-language corpora.

4.3 Experimental results

According to the evaluation scenarios of our experiment listed above, we received the following results:

1. The results of monolingual training and testing of MUSE on English, Hebrew, and Arabic corpora are demonstrated in Tables 4 and 5 for ROUGE-1 and ROUGE-2, respectively. The average ROUGE values obtained using 10-fold cross validation are reported.
2. Tables 4 and 5 show the comparative results for MUSE and unsupervised methods on each corpora, for ROUGE-1 and ROUGE-2 respectively. From Tables 4 and 5 it can be concluded that MUSE performs significantly better (see the statistical analysis of Test 1 below) than other (unsupervised) summarizers in all three corpora,[25] except the baseline in Arabic that was non-distinguishable from MUSE based on ROUGE-2 score (see the explanation of this phenomenon below). According to the *p* values of Test 1, the null hypothesis ("MUSE does not outperform other approaches.") can be rejected at the 0.01 significance level.
3. The results of cross-lingual training are presented in Tables 6a and 7a. From Tables 6a and 7a it can be seen that the null hypothesis of Test 2 ("Training MUSE on source-language corpora does not decrease the summarization quality.") can be rejected only for the Hebrew and Arabic summarizers in most cases. According to the *p* values of Test 2, it can be rejected at the 0.01 significance level. An exception to this conclusion is the Arabic summarizer using the Hebrew model, where the decrease in both ROUGE scores was not significant. Surprisingly, the English summarizer performs significantly better using foreign models then using models trained on the English corpus. The possible reasons for that outcome include a larger number of annotators per each document in the Hebrew and Arabic corpora and the Gold Standard summaries in English being extracts rather than abstracts. Even when trained on the foreign language, MUSE outperforms TextRank for most cases, as can be seen from Tables 4, 5, 6 and 7. Training MUSE on two source corpora instead of one improved the results of training on a single corpus for Hebrew summarizer only.
4. Tables 6b and 7b present the results of applying the summarization model, trained on documents translated into a target language, to original documents in the same target language. For example, applying the model, which was trained on the English corpus (DUC 2002) translated into Hebrew, to the original Hebrew corpus resulted in 0.518

[25] The MUSE *testing* scores, obtained from ten tests in 10-fold cross validation, were compared to the scores of unsupervised summarizers.

ROUGE-1 Recall score (Table 6b, first row, second column). The quality of summarization after training the summarizer on original and translated data is quite close, though statistically distinguishable in most cases (see the statistical analysis of Test 3 below). The results in Tables 6b and 7b demonstrate a significant improvement in summarization quality when the following translated corpora are used—Arabic to English, Arabic to Hebrew, English to Hebrew, and English to Arabic—for summarizing documents in English, Hebrew and Arabic, respectively. In all other cases the null hypothesis cannot be rejected—no significant improvement has been observed. Translating Hebrew to Arabic even decreased the summarization quality in terms of ROUGE-2 scores. Based on these experimental results, it seems that translation may help to get more accurate models and improve the cross-lingual learning of MUSE, given a high-quality machine translation tool for a source-target pair of languages. We suppose that, since Hebrew is a resource-poor language, machine translation from Hebrew suffers from a low quality. Training MUSE on two source corpora instead of one did not improve the results of training on a single corpus.

5. Applying the estimated MLR model to predict the sentence relevance score for summarizing the same set of documents resulted in the 0.426 ROUGE-1 score that equals the Baseline score and is significantly lower than the MUSE score. The results of a pairwise comparison of weights in the two models show that there is no correlation between the two weighting vectors (Pearson correlation $= -0.172$). A possible reason for a difference between GA and MLR models is that in our experiments GA and MLR used slightly different objective functions. The GA fitness function was the ROUGE score of complete document summaries generated by a candidate solution (a global objective), whereas MLR used the ROUGE scores of single sentences as its objective function (in order to obtain a sentence-ranking model) and compiled final summaries from top-ranking sentences. Since the greedy approach does not necessarily solve global optimization problems (knapsack), MLR performed worse as a global optimizer. The optimization procedures are also different: GA explores simultaneously a diverse population of candidate solutions and strikes a balance between exploration and exploitation, whereas the MLR minimizes the sum of squared residuals. Apparently, the GA approach has an advantage in both aspects.

Tables 4, 5, 6 and 7 demonstrate the results of statistical tests, by marking significantly different scores by stars (**p* value of 0.05, ***p* value of 0.01, and ****p* value of 0.001). Tables 4 and 5 contain the results for Test 1, Tables 6a and 7a mark ROUGE scores obtained by cross-lingual training that are significantly lower than the scores obtained by monolingual training (Test 2), and Tables 6b and 7b conclude comparison results between ROUGE scores obtained by cross-lingual training with translation and without it (Test 3).

It can be seen that the obtained ROUGE scores are very different for the three languages: the lowest values were obtained for English summaries, while the highest ones were obtained in the Arabic corpus. This phenomenon can be explained by different characteristics of the gold standard in each corpus. For example, DUC 2002 corpus (English) contains 2–3 *abstracts* for each document, each one of approximately 100 words, the Hebrew corpus consists of five 100-word *extracts* per document in average, and the Arabic corpus contains exactly three 100-word *extracts* per document. Since all evaluated summarization methods generate extracts, their matching with human-generated extracts in Arabic and Hebrew was higher than with English abstracts. Another limitation of gold standard extracts in the Arabic and Hebrew corpora is that many summaries are compiled

Table 4 Mono-lingual training

	ENGLISH	HEBREW	ARABIC
(a) MUSE. 10-fold cross validation			
Mode\corpus			
Train	0.449	0.523	0.751
Test	**0.447**	**0.522**	**0.745**
(b) Unsupervised approaches			
Method\corpus			
Coverage	**0.442****	0.466***	0.723**
Baseline	0.426***	**0.504*****	**0.740***
TextRank	0.425***	0.432***	0.693***
MS Word	0.310***	0.351***	X

The test values in bold are the ones to which those of other approaches were compared and the bold values of the unsupervised approaches indicate the best scores

Mean ROUGE-1 Recall

Table 5 Mono-lingual training

	ENGLISH	HEBREW	ARABIC
(a) MUSE. 10-fold cross validation			
Mode\corpus			
Train	0.211	0.464	0.588
Test	**0.208**	**0.456**	**0.580**
(b) Unsupervised approaches			
Method\corpus			
Coverage	**0.195*****	0.393***	0.518***
Baseline	0.192***	**0.444*****	**0.577**
TextRank	0.172***	0.343***	0.460***

Mean ROUGE-2 Recall

of initial sentences. It causes the superiority of the single unsupervised method (called “baseline”) which takes initial sentences as a summary in both corpora.

Figure 6 present models learned by MUSE on different monolingual corpora using ROUGE-1 and ROUGE-2, respectively. It is noteworthy that while the optimal values of weights in the linear combination were expected to be nonnegative, the actual results in the trained models included some negative values. Although there is no simple explanation for this outcome, it may be related to a well-known phenomenon from Numerical Analysis called *over-relaxation* (Friedman and Kandel 1994). For example, the Laplace equation $\phi_{xx} + \phi_{yy} = 0$ is iteratively solved over a grid of points as follows: At each grid point let $\phi^{(n)}, \overline{\phi}^{(n)}$ denote the *n*th iteration as calculated from the differential equation and its

Table 6 MUSE. Cross-lingual training

	ENGLISH	HEBREW	ARABIC
(a) Cross-lingual training using source language corpora			
Model\corpus			
ENGLISH	X	0.418***	0.723***
HEBREW	0.460***	X	0.745
ARABIC	0.452*	0.515**	X
ENGLISH + HEBREW	X	X	0.734***
ENGLISH + ARABIC	X	0.520	X
HEBREW + ARABIC	0.461***	X	X
(b) Cross-lingual training using translated source corpora			
Source$\xrightarrow{MT}$target			
ENGLISH	X	0.518***	0.737***
HEBREW	0.462	X	0.742
ARABIC	0.457**	0.522**	X
ENGLISH + HEBREW	X	X	0.737
ENGLISH + ARABIC	X	0.513*	X
HEBREW + ARABIC	0.461	X	X

Mean ROUGE-1 Recall

modified final value, respectively. The final value is chosen as $\omega\phi^{(n)} + (1-\omega)\overline{\phi}^{(n-1)}$. While the sum of the two weights is obviously 1, the *optimal* value of ω, which minimizes the number of iterations needed for convergence, usually satisfies $1 < \omega < 2$ (i.e., the second weight $1-\omega$ is negative) and approaches 2 the finer the grid gets. Though somewhat unexpected, this surprising result can be rigorously proved (Varga 1962). Relative to the summarization problem, overrelaxation means using higher positive weights, i.e. "awards" for "better" features and attaching negative weights, i.e. "penalties" to "worse" features. As it can be seen from the charts, there are features that have a similar behavior across languages (for example, position and coverage features), also there are features that always get high positive (POS_F and POS_B) or high negative (POS_L) weights. Some features are correlated (Litvak 2010; Litvak et al. 2010a). However, this should not affect the performance of our method, which chooses the optimal weights of all features simultaneously.

We performed additional experiments for a deep analysis of the GA behavior on our text summarization problem. First, we checked whether termination of the GA ended with the same solutions over multiple cross validation runs, by calculating cosine similarity between these solutions and their centroid (average) vector. According to our experimental results on the Hebrew corpus, the solutions over multiple cross validation runs are very close to each other with the average cosine similarity = 0.75. Second, in order to indicate whether GA "stuck" in a local optima, we calculated the distribution of the last generation of vectors. According to the experimental results on Hebrew corpus, the last generation of vectors appears to be relatively diverse, since only 20 % of the final population has a cosine similarity of more than 0.5 to the centroid vector. Since high genotypic diversity is supposed to prevent premature convergence to a local optimum (Burke et al. 2004), this

Table 7 MUSE. Cross-lingual training

	ENGLISH	HEBREW	ARABIC
(a) Cross-lingual training using source language corpora			
Model\corpus			
ENGLISH	X	0.330***	0.508***
HEBREW	0.219**	X	0.577
ARABIC	0.213*	0.451*	X
ENGLISH + HEBREW	X	X	0.548***
ENGLISH + ARABIC	X	0.456	X
HEBREW + ARABIC	0.220**	X	X
(b) Cross-lingual training using translated source corpora			
Source$\xrightarrow{MT}$target			
ENGLISH	X	0.455***	0.540***
HEBREW	0.220	X	0.556**
ARABIC	0.217*	0.462***	X
ENGLISH + HEBREW	X	X	0.544
ENGLISH + ARABIC	X	0.451	X
HEBREW + ARABIC	0.220	X	X

Mean ROUGE-2 Recall

may lead us to the conclusion that the best fitness in our final population may actually be close to the global optimum.

Figures 7, 8, and 9 demonstrate sample documents and their summaries—in a source language and translated to English—generated by MUSE for the Arabic, Hebrew and English languages, respectively. The summaries' length was restricted to 100 words. It can be seen that the summaries contain the most informative sentences from the original documents, avoiding small details.

5 Conclusions and future work

In this article, monolingual and cross-lingual methods for training MUSE—a *supervised* approach for *multilingual* summarization were described and evaluated on three different languages: English, Hebrew, and Arabic. The evaluation included three different scenarios: (1) retraining for each new language on a new corpus of documents in the target language, (2) using the same training model across different languages, and (3) using parallel corpora (based on machine translation) for retraining MUSE on each new language.

The experimental results show that MUSE significantly outperforms TextRank, the best known language-independent approach, in three languages and all scenarios using either monolingual or parallel corpora. The results also suggest that the same weighting model is applicable across multiple languages and, despite a statistically distinguishable decrease in the summarization quality compared to the mono-lingual summarization, this approach still preserves a reasonable level of quality while saving the annotation efforts for each target

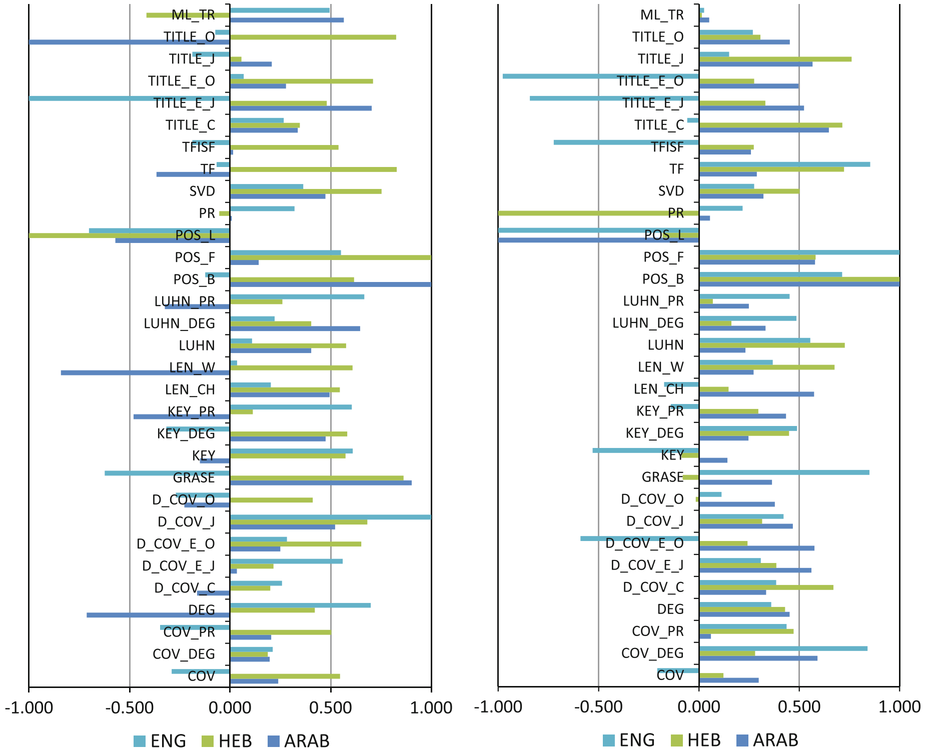


Fig. 6 Models trained on monolingual corpora: ROUGE-1 (left) and ROUGE-2 (right)

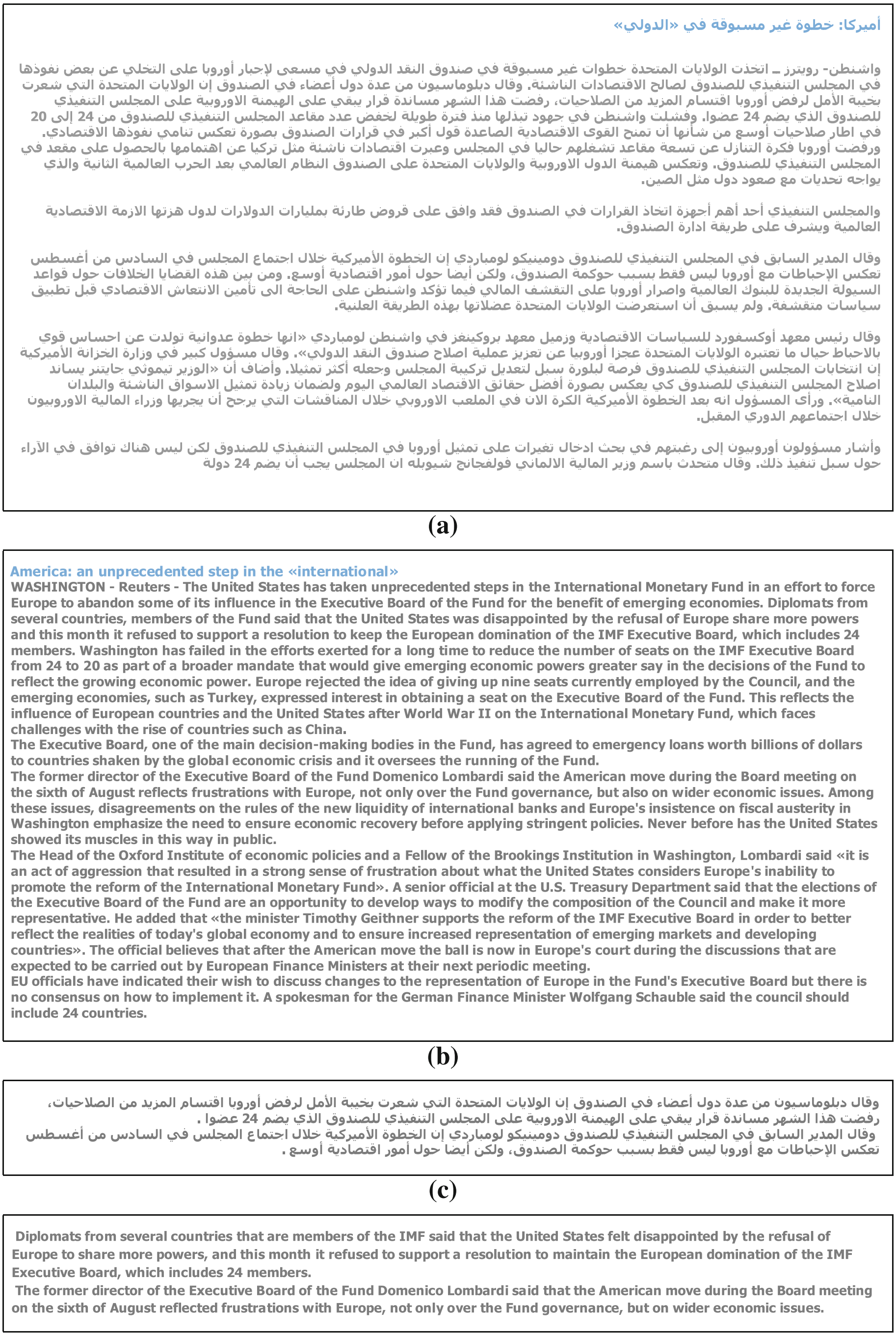
language. On the other hand, using translated corpora may improve the cross-lingual performance of MUSE versus training on source-language corpora, while requiring a minor effort from the end user in preparing the machine-translated version of an existing corpus in any language.

During our research we tried to analyse the reasons of MUSE's superiority by experimenting with different settings:

1. Replace GA by another (MLR) learning procedure (see Sect. 4.3 above),
2. Reduce the number of features (Litvak et al. 2010a),
3. Restrict all feature weights to non-negative values only (Litvak 2010).

According to our results, we can conclude that MUSE has reached its performance superiority due to a large set of features relevant to the summarization task combined with GA as a good choice for optimizing a linear combination of those features. Allowing both positive and negative weights in the linear combination has improved the results as well.

Generally, we can conclude that a combination of as many independent statistical features as possible can compensate for the lack of linguistic analysis and knowledge when selecting the most informative sentences for a summary. One can add more sentence features, and/or use another sophisticated supervised model for learning and optimizing a feature combination. We believe that such approach would work in a general case–retraining on different genres and languages.



أميركا: خطوة غير مسبوقة في «الدولي»

واشنطن- رويترز ـ اتخذت الولايات المتحدة خطوات غير مسبوقة في صندوق النقد الدولي في مسعى لإجبار أوروبا على التخلي عن بعض نفوذها في المجلس التنفيذي للصندوق لصالح الاقتصادات الناشئة. وقال دبلوماسيون من عدة دول أعضاء في الصندوق إن الولايات المتحدة التي شعرت بخيبة الأمل لرفض أوروبا اقتسام المزيد من الصلاحيات، رفضت هذا الشهر مساندة قرار يبقي على الهيمنة الاوروبية على المجلس التنفيذي للصندوق الذي يضم 24 عضوا. وفشلت واشنطن في جهود تبذلها منذ فترة طويلة لخفض عدد مقاعد المجلس التنفيذي للصندوق من 24 إلى 20 في اطار صلاحيات أوسع من شأنها أن تمنح القوى الاقتصادية الصاعدة قول أكبر في قرارات الصندوق بصورة تعكس تنامي نفوذها الاقتصادي. ورفضت أوروبا فكرة التنازل عن تسعة مقاعد تشغلهم حاليا في المجلس وعبرت اقتصادات ناشئة مثل تركيا عن اهتمامها بالحصول على مقعد في المجلس التنفيذي للصندوق. وتعكس هيمنة الدول الاوروبية والولايات المتحدة على الصندوق النظام العالمي بعد الحرب العالمية الثانية والذي يواجه تحديات مع صعود دول مثل الصين.

والمجلس التنفيذي أحد أهم أجهزة اتخاذ القرارات في الصندوق فقد وافق على قروض طارئة بمليارات الدولارات لدول هزتها الازمة الاقتصادية العالمية ويشرف على طريقة ادارة الصندوق.

وقال المدير السابق في المجلس التنفيذي للصندوق دومينيكو لومباردي إن الخطوة الأميركية خلال اجتماع المجلس في السادس من أغسطس تعكس الإحباطات مع أوروبا ليس فقط بسبب حوكمة الصندوق، ولكن أيضا حول أمور اقتصادية أوسع. ومن بين هذه القضايا الخلافات حول قواعد السيولة الجديدة للبنوك العالمية واصرار أوروبا على التقشف المالي فيما تؤكد واشنطن على الحاجة الى تأمين الانتعاش الاقتصادي قبل تطبيق سياسات متقشفة. ولم يسبق أن استعرضت الولايات المتحدة عضلاتها بهذه الطريقة العلنية.

وقال رئيس معهد أوكسفورد للسياسات الاقتصادية وزميل معهد بروكينغز في واشنطن لومباردي «انها خطوة عدوانية تولدت عن احساس قوي بالاحباط حيال ما تعتبره الولايات المتحدة عجزا أوروبيا عن تعزيز عملية اصلاح صندوق النقد الدولي». وقال مسؤول كبير في وزارة الخزانة الأميركية إن انتخابات المجلس التنفيذي للصندوق فرصة لبلورة سبل لتعديل تركيبة المجلس وجعله أكثر تمثيلا. وأضاف أن «الوزير تيموثي جايتنر يساند اصلاح المجلس التنفيذي للصندوق كي يعكس بصورة أفضل حقائق الاقتصاد العالمي اليوم ولضمان زيادة تمثيل الاسواق الناشئة والبلدان النامية». ورأى المسؤول انه بعد الخطوة الأميركية الكرة الان في الملعب الاوروبي خلال المناقشات التي يرجح أن يجريها وزراء المالية الاوروبيون خلال اجتماعهم الدوري المقبل.

وأشار مسؤولون أوروبيون إلى رغبتهم في بحث ادخال تغيرات على تمثيل أوروبا في المجلس التنفيذي للصندوق لكن ليس هناك توافق في الآراء حول سبل تنفيذ ذلك. وقال متحدث باسم وزير المالية الالماني فولفجانج شيوبله ان المجلس يجب أن يضم 24 دولة

(a)

America: an unprecedented step in the «international»

WASHINGTON - Reuters - The United States has taken unprecedented steps in the International Monetary Fund in an effort to force Europe to abandon some of its influence in the Executive Board of the Fund for the benefit of emerging economies. Diplomats from several countries, members of the Fund said that the United States was disappointed by the refusal of Europe share more powers and this month it refused to support a resolution to keep the European domination of the IMF Executive Board, which includes 24 members. Washington has failed in the efforts exerted for a long time to reduce the number of seats on the IMF Executive Board from 24 to 20 as part of a broader mandate that would give emerging economic powers greater say in the decisions of the Fund to reflect the growing economic power. Europe rejected the idea of giving up nine seats currently employed by the Council, and the emerging economies, such as Turkey, expressed interest in obtaining a seat on the Executive Board of the Fund. This reflects the influence of European countries and the United States after World War II on the International Monetary Fund, which faces challenges with the rise of countries such as China.

The Executive Board, one of the main decision-making bodies in the Fund, has agreed to emergency loans worth billions of dollars to countries shaken by the global economic crisis and it oversees the running of the Fund.

The former director of the Executive Board of the Fund Domenico Lombardi said the American move during the Board meeting on the sixth of August reflects frustrations with Europe, not only over the Fund governance, but also on wider economic issues. Among these issues, disagreements on the rules of the new liquidity of international banks and Europe's insistence on fiscal austerity in Washington emphasize the need to ensure economic recovery before applying stringent policies. Never before has the United States showed its muscles in this way in public.

The Head of the Oxford Institute of economic policies and a Fellow of the Brookings Institution in Washington, Lombardi said «it is an act of aggression that resulted in a strong sense of frustration about what the United States considers Europe's inability to promote the reform of the International Monetary Fund». A senior official at the U.S. Treasury Department said that the elections of the Executive Board of the Fund are an opportunity to develop ways to modify the composition of the Council and make it more representative. He added that «the minister Timothy Geithner supports the reform of the IMF Executive Board in order to better reflect the realities of today's global economy and to ensure increased representation of emerging markets and developing countries». The official believes that after the American move the ball is now in Europe's court during the discussions that are expected to be carried out by European Finance Ministers at their next periodic meeting.

EU officials have indicated their wish to discuss changes to the representation of Europe in the Fund's Executive Board but there is no consensus on how to implement it. A spokesman for the German Finance Minister Wolfgang Schauble said the council should include 24 countries.

(b)

وقال دبلوماسيون من عدة دول أعضاء في الصندوق إن الولايات المتحدة التي شعرت بخيبة الأمل لرفض أوروبا اقتسام المزيد من الصلاحيات، رفضت هذا الشهر مساندة قرار يبقي على الهيمنة الاوروبية على المجلس التنفيذي للصندوق الذي يضم 24 عضوا .
وقال المدير السابق في المجلس التنفيذي للصندوق دومينيكو لومباردي إن الخطوة الأميركية خلال اجتماع المجلس في السادس من أغسطس تعكس الإحباطات مع أوروبا ليس فقط بسبب حوكمة الصندوق، ولكن أيضا حول أمور اقتصادية أوسع .

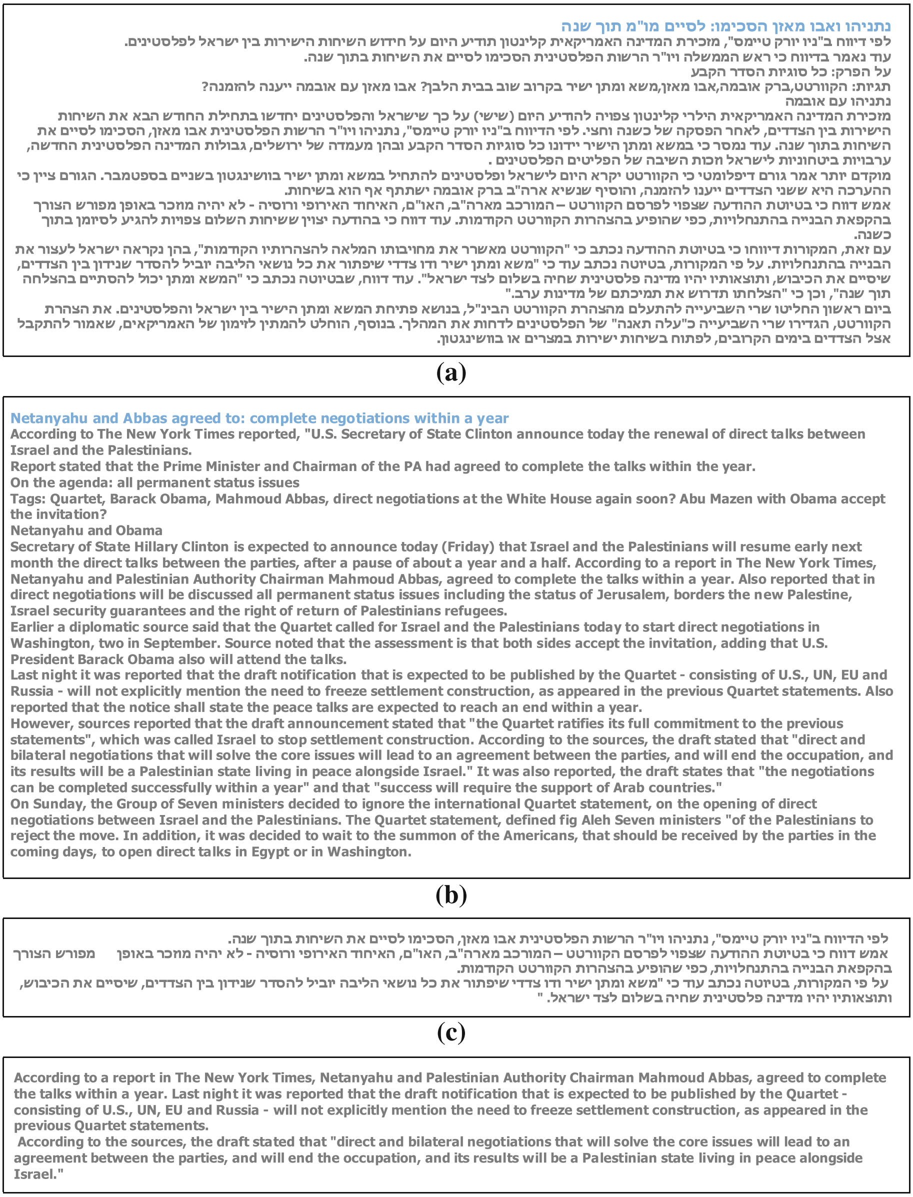
(c)

Diplomats from several countries that are members of the IMF said that the United States felt disappointed by the refusal of Europe to share more powers, and this month it refused to support a resolution to maintain the European domination of the IMF Executive Board, which includes 24 members.

The former director of the Executive Board of the Fund Domenico Lombardi said that the American move during the Board meeting on the sixth of August reflected frustrations with Europe, not only over the Fund governance, but on wider economic issues.

(d)

Fig. 7 Arabic document titled "*America: an unprecedented step in the*"*International (Monetary Fund)*"" and its summary. **a** Source document, **b** translated document , **c** original summary, **d** translated summary



נתניהו ואבו מאזן הסכימו: לסיים מו"מ תוך שנה

לפי דיווח ב"ניו יורק טיימס", מזכירת המדינה האמריקאית קלינטון תודיע היום על חידוש השיחות הישירות בין ישראל לפלסטינים.

עוד נאמר בדיווח כי ראש הממשלה ויו"ר הרשות הפלסטינית הסכימו לסיים את השיחות בתוך שנה.

על הפרק: כל סוגיות הסדר הקבע

תגיות: הקוורטט,ברק אובמה,אבו מאזן,משא ומתן ישיר בקרוב שוב בבית הלבן? אבו מאזן עם אובמה ייענה להזמנה?

נתניהו עם אובמה

מזכירת המדינה האמריקאית הילרי קלינטון צפויה להודיע היום (שישי) על כך שישראל והפלסטינים יחדשו בתחילת החודש הבא את השיחות הישירות בין הצדדים, לאחר הפסקה של כשנה וחצי. לפי הדיווח ב"ניו יורק טיימס", נתניהו ויו"ר הרשות הפלסטינית אבו מאזן, הסכימו לסיים את השיחות בתוך שנה. עוד נמסר כי במשא ומתן הישיר יידונו כל סוגיות הסדר הקבע ובהן מעמדה של ירושלים, גבולות המדינה הפלסטינית החדשה, ערבויות ביטחוניות לישראל וזכות השיבה של הפליטים הפלסטינים .

מוקדם יותר אמר גורם דיפלומטי כי הקוורטט יקרא היום לישראל ופלסטינים להתחיל במשא ומתן ישיר בוושינגטון בשניים בספטמבר. הגורם ציין כי ההערכה היא ששני הצדדים ייענו להזמנה, והוסיף שנשיא ארה"ב ברק אובמה ישתתף אף הוא בשיחות.

אמש דווח כי בטיוטת ההודעה שצפוי לפרסם הקוורטט – המורכב מארה"ב, האו"ם, האיחוד האירופי ורוסיה - לא יהיה מוזכר באופן מפורש הצורך בהקפאת הבנייה בהתנחלויות, כפי שהופיע בהצהרות הקוורטט הקודמות. עוד דווח כי בהודעה יצוין ששיחות השלום צפויות להגיע לסיומן בתוך כשנה.

עם זאת, המקורות דיווחו כי בטיוטת ההודעה נכתב כי "הקוורטט מאשרר את מחויבותו המלאה להצהרותיו הקודמות", בהן נקראה ישראל לעצור את הבנייה בהתנחלויות. על פי המקורות, בטיוטה נכתב עוד כי "משא ומתן ישיר ודו צדדי שיפתור את כל נושאי הליבה יוביל להסדר שנידון בין הצדדים, שיסיים את הכיבוש, ותוצאותיו יהיו מדינה פלסטינית שחיה בשלום לצד ישראל". עוד דווח, שבטיוטה נכתב כי "המשא ומתן יכול להסתיים בהצלחה תוך שנה", וכן כי "הצלחתו תדרוש את תמיכתם של מדינות ערב."

ביום ראשון החליטו שרי השביעייה להתעלם מהצהרת הקוורטט הבינ"ל, בנושא פתיחת המשא ומתן הישיר בין ישראל והפלסטינים. את הצהרת הקוורטט, הגדירו שרי השביעייה כ"עלה תאנה" של הפלסטינים לדחות את המהלך. בנוסף, הוחלט להמתין לזימון של האמריקאים, שאמור להתקבל אצל הצדדים בימים הקרובים, לפתוח בשיחות ישירות במצרים או בוושינגטון.

(a)

Netanyahu and Abbas agreed to: complete negotiations within a year

According to The New York Times reported, "U.S. Secretary of State Clinton announce today the renewal of direct talks between Israel and the Palestinians.

Report stated that the Prime Minister and Chairman of the PA had agreed to complete the talks within the year.

On the agenda: all permanent status issues

Tags: Quartet, Barack Obama, Mahmoud Abbas, direct negotiations at the White House again soon? Abu Mazen with Obama accept the invitation?

Netanyahu and Obama

Secretary of State Hillary Clinton is expected to announce today (Friday) that Israel and the Palestinians will resume early next month the direct talks between the parties, after a pause of about a year and a half. According to a report in The New York Times, Netanyahu and Palestinian Authority Chairman Mahmoud Abbas, agreed to complete the talks within a year. Also reported that in direct negotiations will be discussed all permanent status issues including the status of Jerusalem, borders the new Palestine, Israel security guarantees and the right of return of Palestinians refugees.

Earlier a diplomatic source said that the Quartet called for Israel and the Palestinians today to start direct negotiations in Washington, two in September. Source noted that the assessment is that both sides accept the invitation, adding that U.S. President Barack Obama also will attend the talks.

Last night it was reported that the draft notification that is expected to be published by the Quartet - consisting of U.S., UN, EU and Russia - will not explicitly mention the need to freeze settlement construction, as appeared in the previous Quartet statements. Also reported that the notice shall state the peace talks are expected to reach an end within a year.

However, sources reported that the draft announcement stated that "the Quartet ratifies its full commitment to the previous statements", which was called Israel to stop settlement construction. According to the sources, the draft stated that "direct and bilateral negotiations that will solve the core issues will lead to an agreement between the parties, and will end the occupation, and its results will be a Palestinian state living in peace alongside Israel." It was also reported, the draft states that "the negotiations can be completed successfully within a year" and that "success will require the support of Arab countries."

On Sunday, the Group of Seven ministers decided to ignore the international Quartet statement, on the opening of direct negotiations between Israel and the Palestinians. The Quartet statement, defined fig Aleh Seven ministers "of the Palestinians to reject the move. In addition, it was decided to wait to the summon of the Americans, that should be received by the parties in the coming days, to open direct talks in Egypt or in Washington.

(b)

לפי הדיווח ב"ניו יורק טיימס", נתניהו ויו"ר הרשות הפלסטינית אבו מאזן, הסכימו לסיים את השיחות בתוך שנה.

אמש דווח כי בטיוטת ההודעה שצפוי לפרסם הקוורטט – המורכב מארה"ב, האו"ם, האיחוד האירופי ורוסיה - לא יהיה מוזכר באופן מפורש הצורך בהקפאת הבנייה בהתנחלויות, כפי שהופיע בהצהרות הקוורטט הקודמות.

על פי המקורות, בטיוטה נכתב עוד כי "משא ומתן ישיר ודו צדדי שיפתור את כל נושאי הליבה יוביל להסדר שנידון בין הצדדים, שיסיים את הכיבוש, ותוצאותיו יהיו מדינה פלסטינית שחיה בשלום לצד ישראל. "

(c)

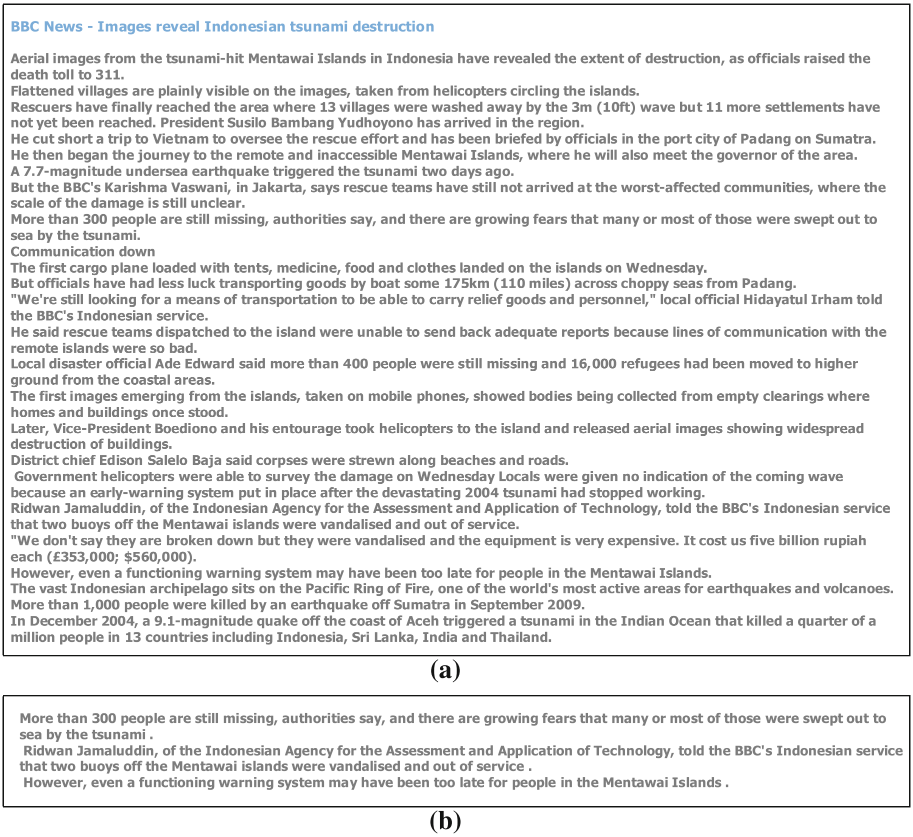
According to a report in The New York Times, Netanyahu and Palestinian Authority Chairman Mahmoud Abbas, agreed to complete the talks within a year. Last night it was reported that the draft notification that is expected to be published by the Quartet - consisting of U.S., UN, EU and Russia - will not explicitly mention the need to freeze settlement construction, as appeared in the previous Quartet statements.

According to the sources, the draft stated that "direct and bilateral negotiations that will solve the core issues will lead to an agreement between the parties, and will end the occupation, and its results will be a Palestinian state living in peace alongside Israel."

(d)

Fig. 8 Hebrew document titled *"Netanyahu and Abbas agreed to complete negotiations within a year"* and its summary. **a** Source document, **b** translated document , **c** original summary, **d** translated summary

Based on evaluation results, the following may be recommended: If a corpus in the target language exists, the best approach is to train MUSE on the target-language corpus, while periodically updating the trained model when new annotated data becomes available. If there is a corpus in a source language, but no high-quality target-language corpus is



BBC News - Images reveal Indonesian tsunami destruction

Aerial images from the tsunami-hit Mentawai Islands in Indonesia have revealed the extent of destruction, as officials raised the death toll to 311.
Flattened villages are plainly visible on the images, taken from helicopters circling the islands.
Rescuers have finally reached the area where 13 villages were washed away by the 3m (10ft) wave but 11 more settlements have not yet been reached. President Susilo Bambang Yudhoyono has arrived in the region.
He cut short a trip to Vietnam to oversee the rescue effort and has been briefed by officials in the port city of Padang on Sumatra.
He then began the journey to the remote and inaccessible Mentawai Islands, where he will also meet the governor of the area.
A 7.7-magnitude undersea earthquake triggered the tsunami two days ago.
But the BBC's Karishma Vaswani, in Jakarta, says rescue teams have still not arrived at the worst-affected communities, where the scale of the damage is still unclear.
More than 300 people are still missing, authorities say, and there are growing fears that many or most of those were swept out to sea by the tsunami.
Communication down
The first cargo plane loaded with tents, medicine, food and clothes landed on the islands on Wednesday.
But officials have had less luck transporting goods by boat some 175km (110 miles) across choppy seas from Padang.
"We're still looking for a means of transportation to be able to carry relief goods and personnel," local official Hidayatul Irham told the BBC's Indonesian service.
He said rescue teams dispatched to the island were unable to send back adequate reports because lines of communication with the remote islands were so bad.
Local disaster official Ade Edward said more than 400 people were still missing and 16,000 refugees had been moved to higher ground from the coastal areas.
The first images emerging from the islands, taken on mobile phones, showed bodies being collected from empty clearings where homes and buildings once stood.
Later, Vice-President Boediono and his entourage took helicopters to the island and released aerial images showing widespread destruction of buildings.
District chief Edison Salelo Baja said corpses were strewn along beaches and roads.
Government helicopters were able to survey the damage on Wednesday Locals were given no indication of the coming wave because an early-warning system put in place after the devastating 2004 tsunami had stopped working.
Ridwan Jamaluddin, of the Indonesian Agency for the Assessment and Application of Technology, told the BBC's Indonesian service that two buoys off the Mentawai islands were vandalised and out of service.
"We don't say they are broken down but they were vandalised and the equipment is very expensive. It cost us five billion rupiah each (£353,000; $560,000).
However, even a functioning warning system may have been too late for people in the Mentawai Islands.
The vast Indonesian archipelago sits on the Pacific Ring of Fire, one of the world's most active areas for earthquakes and volcanoes.
More than 1,000 people were killed by an earthquake off Sumatra in September 2009.
In December 2004, a 9.1-magnitude quake off the coast of Aceh triggered a tsunami in the Indian Ocean that killed a quarter of a million people in 13 countries including Indonesia, Sri Lanka, India and Thailand.

(a)

More than 300 people are still missing, authorities say, and there are growing fears that many or most of those were swept out to sea by the tsunami .
Ridwan Jamaluddin, of the Indonesian Agency for the Assessment and Application of Technology, told the BBC's Indonesian service that two buoys off the Mentawai islands were vandalised and out of service .
However, even a functioning warning system may have been too late for people in the Mentawai Islands .

(b)

Fig. 9 English document titled *"Images reveal Indonesian tsunami destruction"* and its summary. **a** Source document, **b** summary

available, the recommendation is to create a machine-translated corpus for the target language and apply cross-lingual learning of MUSE using this parallel corpora. Using any *unsupervised* method which does not require training on any language is not recommended, since none of these methods were found to outperform MUSE on any of the three languages.

In the future work, it is suggested to evaluate MUSE on additional languages, language families, and genres, incorporate threshold values for threshold-based methods (Table 2) into the GA-based optimization procedure, improve performance of similarity-based methods in the multilingual domain, apply additional optimization techniques like Evolution Strategy (Beyer and Schwefel 2002), which is known to perform well in a real-valued search space, reduce the search for the best summary to the problem of multi-objective optimization, combining several summary quality metrics, extend the Arabic and Hebrew corpora to improve the quality of the trained summarization model, and adapt the MUSE approach to multi-document summarization.

Acknowledgments We are grateful to Michael Elhadad and Galina Volk from the Ben-Gurion University for providing the ROUGE toolkit adapted to the Hebrew alphabet, and to Slava Kisilevich from the University of Konstanz for the technical support in evaluation experiments. Also, we would like to thank Michael Orlov and Menahem Friedman for consulting on genetic algorithms.

References

Aker, A., Cohn, T., & Gaizauskas, R. (2010). Multi-document summarization using A* search and discriminative training. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 482–491). Cambridge, Massachusetts.

Alfonseca, E., & Rodriguez, P. (2003). Generating extracts with genetic algorithms. In *Proceedings of the 2003 European conference on information retrieval (ECIR'2003)* (pp. 511–519). Pisa, Italy.

Baxendale, P. B. (1958). Machine-made index for technical literaturean experiment. *IBM Journal of Research and Development, 2*, 354–361.

Beyer, H.-G., & Schwefel, H.-P. (2002) Evolution strategies: A comprehensive introduction. *Natural Computing, 1*, 3–52.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems, 30*, 107–117.

Burke, E., Gustafson, S., & Kendall, G. (2004). Diversity in genetic programming: An analysis of measures and correlation with fitness. *IEEE Transactions on Evolutionary Computation, 8*, 47–62.

DUC (2002). Document understanding conference. http://duc.nist.gov.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM 16*(2), 264–285.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22*, 457–479.

Friedman, M., & Kandel, A. (1994). *Fundamentals of computer numerical analysis*. Boca Raton, FL: CRC Press.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Reading, MA: Addison-Wesley.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 121–128). Berkeley, California, USA.

Gulli, A., & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. http://www.cs.uiowa.edu/~asignori/web-size/.

Hassel, M., & Sjobergh, J. (2006). Towards holistic summarization: Selecting summaries, not sentences. In *Proceedings of LREC—International conference on language resources and evaluation*.

Hovy, E. (2001). *Multilingual information management: Current levels and future abilities*. Linguistica computazionale. Istituti editoriali e poligrafici internazionali.

Kallel, F. J., Jaoua, M., Hadrich, L. B., & Hamadou, A. B. (2004). Summarization at LARIS Laboratory. In *Proceedings of the document understanding conference*.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM), 46*, 604–632.

Lin, C., & Hovy, E. (1997). Identifying topics by position. In *Proceedings of the fifth conference on applied natural language processing* (pp. 283–290).

Lin, C.-Y. (2004a). Looking for a few good metrics: Automatic summarization evaluation—How many samples are enough?. In *Proceedings of NTCIR-4* (pp. 1–10).

Lin, C.-Y. (2004b). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (WAS 2004)* (pp. 25–26).

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *NAACL '03: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology* (pp. 71–78). Edmonton, Canada.

Litvak, M. (2010). *New methodologies for language-independent extractive summarization*. Doctoral dissertation, Kreitman School of Advanced Graduate Studies.

Litvak, M., Aizenman, H., Gobits, I., Last, M., & Kandel, A. (2011). DegExt—A language-independent graph-based keyphrase extractor. In *Proceedings of the 7th Atlantic web intelligence conference (AWIC 2011)* (pp. 121–130). Fribourg, Switzerland.

Litvak, M., Kisilevich, S., Keim, D., Lipman, H., Gur, A. B., & Last, M. (2010a). Towards language-independent summarization: A comparative analysis of sentence extraction methods on English and Hebrew corpora. In *Proceedings of the 4th workshop on cross lingual information access (CLIA)* (pp. 61–69). Beijing, China: Coling 2010 Organizing Committee.

Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the workshop on multi-source multilingual information extraction and summarization* (pp. 17–24).

Litvak, M., Last, M., & Friedman, M. (2010b). A new approach to improving multilingual summarization using a Genetic Algorithm. In *ACL '10: Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 927–936). Uppsala, Sweden.

Liu, D., He, Y., Ji, D., & Yang, H. (2006a). Genetic algorithm based multi-document summarization. In *Proceedings of the 9th Pacific rim international conference on artificial intelligence* (pp. 1140–1144). Guilin, China.

Liu, D., Wang, Y., Liu, C., & Wang, Z. (2006b). Multiple documents summarization based on genetic algorithm. In *Fuzzy systems and knowledge discovery*, Vol. 4223 of *Lecture Notes in Computer Science* (pp. 355–364).

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development, 2*, 159–165.

Mani, I. (2001). *Automatic summarization. Natural language processing*. Amsterdam: John Benjamins Publishing Company.

Mihalcea R. (2005). Language independent extractive summarization. In *AAAI'05: Proceedings of the 20th national conference on artificial intelligence* (pp. 1688–1689).

Neto, J., Santos, A., Kaestner, C., & Freitas, A. (2000). Generating text summaries through the relative importance of topics. *Lecture Notes in Computer Science* (pp. 300–309).

Nobata, C., Sekine, S., Murata, M., Uchimoto, K., Utiyama, M., & Isahara, H. (2001). Sentence extraction system assembling multiple evidence. In *Proceedings of 2nd NTCIR workshop* (pp. 319–324).

Orăsan, C., Evans, R., & Mitkov, R. (2000). Enhancing preference-based anaphora resolution with genetic algorithms. In *Proceedings of the second international conference on natural language processing* (pp. 185–195). Patras, Greece.

Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing and Management, 47*, 227–237.

Radev, D., Blair-Goldensohn, S., & Zhang, Z. (2001). Experiments in single and multidocument summarization using MEAD. In *Proceedings of the first document understanding conference (DUC).*

Saggion, H., Bontcheva, K., & Cunningham, H. (2003). Robust generic and query-based summarisation. In *EACL '03: Proceedings of the tenth conference on European chapter of the association for computational linguistics.*

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing and Management, 33*, 193–207.

Schenker, A., Bunke, H., Last, M., & Kandel, A. (2004). Classification of web documents using graph matching. *International Journal of Pattern Recognition and Artificial Intelligence, 18*, 475–496.

Schenker, A., Bunke, H., Last, M., & Kandel, A. (2005). *Graph-theoretic techniques for web content mining*. Singapore: World Scientific

Steinberger, J., & Jezek, K. (2004). Text summarization and singular value decomposition. *Lecture Notes in Computer Science*, pp. 245–254.

Teufel, S., & Moens, M. (1997). Sentence extraction as a classification task. In *Proceedings of the workshop on intelligent scalable summarization, ACL/EACL conference* (pp. 58–65). Madrid, Spain.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval, 2*, 303–336.

Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management, 43*, 1606–1618.

Varga, R. (1962). *Matrix iterative methods*. Englewood Cliffs NJ: Prentice-Hall.

Vignaux, G. A., & Michalewicz, Z. (1991). A genetic algorithm for the linear transportation problem. *IEEE Transactions on Systems, Man and Cybernetics, 21*, 445–452.

Wan, X. (2008). Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval, 11*, 25–49.

Wan, X., Li, H., & Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *ACL '10: Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 917–926). Uppsala, Sweden.

Wong, K., Wu, M., & Li, W. (2008). Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd international conference on computational linguistics-Volume 1* (pp. 985–992).