MUSE – A Multilingual Sentence Extractor

Marina Litvak* (corresponding author), Mark Last*, Menahem Friedman* and Slava Kisilevich†

*Ben Gurion University, Beer Sheva, Israel
Email: {litvakm, mlast, fmenahem}@bgu.ac.il
†University of Konstanz, Germany
Email: slaks@dbvis.inf.uni-konstanz.de

Abstract—The MUltilingual Sentence Extractor (MUSE) tool is aimed at multilingual single-document summarization defined by [1] as “processing several languages, with summary in the same language as input”. MUSE consists of two main modules: the training module and the summarization module. The training module is provided with a corpus of summarized texts in one or several languages and it learns the best linear combination of user-specified sentence ranking metrics applying a genetic algorithm to the training data. The summarization module performs real-time sentence extraction by computing sentence ranks according to the weighted model induced in the training phase. MUSE was evaluated on two languages from two different language families - English and Hebrew using the ROUGE-1 Recall measure. MUSE has significantly outperformed the best known state-of-the-art extractive summarization methods and tools in both languages. We have also shown that the summarization module can be successfully applied to languages, which were not included in the training corpus.

Keywords—automated summarization; multi-lingual summarization; genetic algorithm;
Submission category: Demo
Equipment we will bring: laptop
Equipment needed: table, power socket.
Special requirements: none.
Present state: We are integrating the system modules under a common user interface.
Estimated conclusion date: Aug. 31, 2010.
Download video presentation: presentation-Demo

I. 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 information overload, many professionals in a variety of fields must contend with on a daily basis, assist in the automated classification and filtering of documents, and increase search engines precision.

The publication of textual information on the Internet in an ever-growing variety of languages increases the importance of developing multilingual summarization systems. There is a particular need for language-independent statistical tools that can be readily applied to text in any language without depending on labor-intensive development of language-specific linguistic modules.

Since pure statistical methods usually compute a single sentence feature, various attempts have been made to use a combination of several metrics as a ranking function [2], [3]. MUSE implements the language-independent summarization methodology presented by us in [4], which extends this effort by learning the best linear combination of 31 statistical language-independent sentence ranking features using a Genetic Algorithm (GA). The sentence features comprising the linear combination are based on several vector and graph representations of text documents requiring a mere word and sentence segmentation of a summarized text written in any language.

The empirical evaluation of MUSE on two monolingual and one bilingual corpora of English and Hebrew documents [4] has shown the following:
- MUSE performance is significantly better than TextRank [5] and Microsoft Word’s Autosummarize tool in both languages.
- In English, MUSE outperforms such known summarization tools as MEAD [2] and SUMMA [3].
- MUSE does not need to be retrained on each language and the same model can be used across at least two different languages.
- MUSE does not need to be trained on a monolingual corpus.

II. MULTILINGUAL SENTENCE EXTRACTOR (MUSE): OVERVIEW

A. Architecture

MUSE is a multilingual extractive single-document summarizer recently developed to summarize documents in different languages by ranking and extracting the most important sentences. The reader is referred to [4] for a detailed description of the multilingual summarization methodology implemented by MUSE. The current version of MUSE tool can be applied only to text documents or textual content of HTML pages. It consists of two main modules: the training module activated in offline and the real-time summarization module, which are shown in the left and the right parts of Figure 1 respectively. Both modules utilize several vector- and graph-based document representations (described in [4]). The preprocessing module is responsible
for document parsing and representation, and it is integrated with both modules.

The training module receives as input a corpus of documents, each accompanied by one or several gold-standard summaries—abstracts or extracts—compiled by human 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. The training module computes a set of user-specified statistical features for each sentence in every document and then applies a genetic algorithm to a document-feature matrix of precomputed sentence scores with the purpose of finding the best linear combination of sentence ranking features. The current version of MUSE is using ROUGE-1 Recall metric as a fitness function though additional summarization quality metrics can be easily added to the system. The output/model of the training module is a vector of weights for the user-specified sentence ranking features. In the current version of the tool, we have implemented 31 vector-based and graph-based features.

The summarization module performs on-line summarization of input text/texts in any language. No manual or automatic language identification is needed prior to summarizing a document. Each sentence of an input text obtains a relevance score according to the trained model, and the top ranked sentences are extracted to the summary in their original order. The length of resulting summaries is limited by a user-specified value (maximum number of words / sentences in the text extract or a length ratio). Being activated in real-time, the summarization module is required 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), When activating the summarization module on a document or a set of documents, the user can choose one of the following options: (1) use the model trained on the same language as input texts, (2) use the the model trained on some other language/corpus (in we show that the same model can be efficiently used across different languages), (3) apply user-specified weights for each method in the combination , and (4) provide user-specified individual sentence scoring method/methods.

The preprocessing module performs the following tasks: (1) sentence segmentation, (2) word segmentation, (3) vector space model construction using tf and/or tf-idf weights, (4) a word-based graph representation construction, and (5) document metadata construction. The outputs of this submodule are: sentence segmented text (SST), vector space model (VSM), and the document graph. Steps (1) and (2) are performed by the text processor submodule, which is implemented using Strategy Design Pattern and consists of three elements: filter, reader and sentence segmenter. The filter works on the Unicode character level and performs such operations as identification of characters, digits, punctuations or normalization (if available for specific source languages). The reader invokes the filter, constructs word chunks from the input stream and identifies the following states: words, special characters, white spaces, numbers, URL links and punctuation marks. The sentence segmenter invokes the Reader and divides the input space into sentences. By implementing different filters, the Reader can work either with a specific language (taking into account its intricacies) or with documents written in arbitrary language. Figure 2 presents a small text example and its graph representation, visual and xml.

Figure 1 shows the general architecture of the MUSE system. Figure 4 in Appendix, shows extract produced by MUSE for one of the DUC 2002 documents.

B. Use Cases

MUSE has three possible use cases demonstrated in Figure 3 and briefly described in Table I settings specification, training, and summarization.

In the settings specification, the user is required to specify the following parameters: the paths to the input documents being summarized, the gold standard, and the output summaries, as well as the maximal length of a summary. An advanced user can change the default settings for a GA: population size, crossover and mutation probabilities, and the minimum improvement parameter. The user can also modify the training settings (splitting into the training and the testing documents) and the preprocessing settings: maximal size of a graph representation, skip numbers, use a list of stopwords (if available), etc. All specified settings can be used in the subsequent training and/or summarization operations or stored for the later use.

C. System Features

MUSE software has the following key features:

- Multilingual Analysis. The way in which MUSE processes a text is fully multi-lingual. All statistical metrics for sentence ranking used by MUSE do not require any language-specific analysis or knowledge, allowing MUSE to process texts in any language even without identifying the language of a given document.
### Table I

#### USE CASE DESCRIPTION

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Goal</th>
<th>Precondition</th>
<th>Postcondition</th>
<th>Brief</th>
</tr>
</thead>
<tbody>
<tr>
<td>Settings</td>
<td>Specify</td>
<td>None</td>
<td>Stored configuration</td>
<td>User specifies all necessary parameters as: path to input document/s, summary length, gold standard folder, etc.</td>
</tr>
<tr>
<td>Specification</td>
<td>summarizer settings</td>
<td></td>
<td></td>
<td>User can store his settings for later use.</td>
</tr>
<tr>
<td>Summarization</td>
<td>Summarize</td>
<td>Settings and</td>
<td>Summary</td>
<td>User can get a summary for the input document/s and store it.</td>
</tr>
</tbody>
</table>

Hurricane Gilbert Heads Toward Dominican Coast.

Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.

The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.

---

- **Optional integration of language-specific tools and features.** The user can add the following language-specific steps to the MUSE pre-processing module: stopwords removal, stemming, sentence segmentation, and POS tagging, by configuring the system and providing the necessary tools or data.

- **Easy to use.** Summarize an entire folder of text or HTML documents with just one click. The resulting summaries can be saved in plain text or HTML format.

- **Output interpretability.** MUSE automatically highlights the sentences that are expected to cover the main aspects of the document’s content.

- **Multiple models.** MUSE allows to store and use multiple ranking models trained on annotated corpora in multiple languages.

- **No need in retraining on every new language.** MUSE allows to use the same trained model across different languages.

- **Two-level configuration.** Typical users can work in the default mode whereas advanced users can configure the settings of a genetic algorithm, preprocessing, etc.

- **Metric flexibility.** The user can choose any subset of sentence metrics to be included in the ranking model.

- **Output flexibility.** The user determines the number of sentences to be included in the summary using the following criteria: maximum number of words, maximum number of sentences or maximum summary-to-document length ratio.

---

### III. CONCLUSIONS

In this demo, we have presented the MUltilingual Sentence Extractor (MUSE) tool aimed at automated summarization of text documents in any language. The tool implements the language-independent single-document summarization methodology introduced and evaluated by us in [4]. It is based on a statistical sentence scoring model learned from a training corpus of summarized documents using a genetic algorithm. The current version of the MUSE tool has several unique features such as summarizing texts in arbitrary languages, cross-lingual use of multiple ranking
models, a rich choice of statistical sentence features, flexible pre-processing and optimization options, etc.

The tool can be enhanced in the future to allow the use of additional optimization techniques, nonlinear scoring models, and multiple summary quality metrics. It can also be integrated with language-specific NLP tools and provide summaries with key phrases rather than key sentences. Data mining and text mining researchers will be able to use this tool for extensive experimentation with document corpora in multiple languages and genres.

REFERENCES


APPENDIX

Figure 4. Input file (AP880228-0097 from DUC 2002 collection) and its extract by MUSE.