MUSEEC: A Multilingual Text Summarization Tool

Marina Litvak[1], Natalia Vanetik[1], Mark Last[2] and Elena Churkin[1]
[1]Department of Software Engineering, Shamoon College of Engineering, Beer Sheva, Israel
[2]Department of Information Systems Engineering, Ben Gurion University of the Negev, Beer Sheva, Israel

The MUSEEC (MUltilingual SEntence Extraction and Compression) summarization tool implements several extractive summarization techniques – at the level of complete and compressed sentences – that can be applied, with some adaptations, to documents in multiple languages. MUSEEC currently supports three languages (English, Hebrew and Arabic) and provides the following summarization methods:

1. MUSE (MUltilingual Sentence Extractor) (Last and Litvak, 2012) – a supervised summarizer, based on a genetic algorithm (GA), that ranks document sentences and extracts top–ranking sentences into a summary. MUSE implements a supervised learning approach to extractive summarization, where the best set of weights for a linear combination of sentence scoring metrics is found by a GA trained on a collection of documents and their gold standard summaries. MUSE training can be performed from the MUSEEC tool. The obtained weighting vector is used for sentence scoring in future summarizations. In the MultiLing 2015 competition of multilingual summarization tools, MUSE has outperformed all other participating systems on English and Arabic corpora, and all systems, except one, on the Hebrew corpus in the single-document summarization task. In the multi-document summarization task, on the other hand, MUSE has outperformed all systems on the Hebrew corpus, and all systems, except one, on English and Arabic corpora. The user can choose a subset of sentence metrics that will be included by MUSE in the linear combination. By default, MUSEEC will use the 31 language-independent metrics presented in (Last and Litvak, 2012). MUSEEC also allows the user to employ additional, linguistic features, which are currently available only for the English language. These features are based on lemmatization, multi-word expressions (MWE), named entity recognition (NER), and Part-of-Speech tagging (POS), all performed with Stanford CoreNLP package for English.

2. POLY (POLYnomial summarizer using POLYtopes) (Litvak and Vanetik, 2013) – an unsupervised summarizer, based on linear programming (LP), that selects the best extract of document sentences. Following the maximum coverage principle, the goal of POLY is to find the best subset of sentences that, under length constraints, can be presented as a summary. It is obvious that the number of potential extracts is exponential in the number of sentences. Therefore, POLY uses an efficient text representation model with the purpose of representing all possible extracts without computing them explicitly. This approach saves a great portion of computation time. Each sentence is represented by a hyperplane, and all sentences derived from a document form hyperplane intersections (polytope). Then, all possible extracts can be represented by subplanes of hyperplane intersections that are located close to the boundary of the polytope. Intuitively, the boundary of the resulting polytope is a good approximation for extracts that can be generated from a given document.

3. WECOM (WEighted COMpression) (Vanetik et al., 2016) – an unsupervised summarizer that compiles a document summary from compressed sentences. We propose to shorten sentences by iteratively removing Elementary Discourse Units – EDUs (grammatically independent parts of a sentence (Marcu, 1997)). We preserve the important content by optimizing the weighting function that measures cumulative importance and preserves a valid syntactic structure of a sentence. The implemented approach consists of the following steps: weight assignment to each term occurrence in the document, preparing and ranking the list of candidate EDUs for removal, and iteratively removing the least significant EDUs from the summary, subject to the summary length constraint.

Tables 1, 2, and 3 contain the summarized results of automated evaluations for the MultiLing 2015, single-document summarization (MSS) task. The quality of the summaries is measured by ROUGE-1 (Recall, Precision, and F-measure), (C.-Y, 2004). We also demonstrate the absolute ranks of each submission–P-Rank, R-Rank, and F-Rank–with their scores sorted by Precision, Recall, and F-measure, respectively. Only the best submissions (in terms of F-measure) for each participating system are presented and sorted in descending order of their F-measure scores. As can be seen, MUSE outperformed all other participating systems except for CCS in Hebrew. MUSEEC also participated in the multidocument summarization (MMS) task, on English, Hebrew and Arabic. MUSE got first place on Hebrew, and 2nd places on English and Arabic languages, out of 9 participants. POLY got third place on Hebrew, 4th place on English, and 5th place on Arabic languages, out of 9 participants. Table 4 contains results for POLY and WECOM summarizers on the DUC 2002 dataset. Statistical testing (using a paired T-test) showed that there is a significant improvement in ROUGE-1 recall between ILP concept-based extraction method of Gillick and Favre (2009) and WECOM with weights generated by Gillick and Favre's method. Another significant improvement is between ILP extraction method of McDonald (2007) and WECOM with weights generated by McDonald's method.

References

M. Last and M. Litvak. 2012. Cross-lingual training of summarization systems using annotated corpora in a foreign language. Information Retrieval, pages 1–28, September.

M. Litvak and N. Vanetik. 2013. Mining the gaps: Towards polynomial summarization. In Proceedings of the International Joint Conference on Natural Language Processing, pages 655–660.

N. Vanetik, M. Litvak, M. Last, and E. Churkin. 2016. An unsupervised constrained optimization approach to compressive summarization. Manuscript submitted for publication.

D. Marcu. 1997. From discourse structures to text summaries. In Proceedings of the ACL, volume 97, pages 82–88.

Lin C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), pages 25–26.

D. Gillick and B. Favre. 2009. A scalable global model for summarization. In Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing.

R. McDonald. 2007. A study of global inference algorithms in multi-document summarization. In Advances in Information Retrieval, pages 557–564.

Figures

system	P score	R score	F score	P-Rank	R-Rank	F-Rank
Oracles	0.601	0.619	0.610	1	1	1
MUSE	**0.488**	**0.500**	**0.494**	**2**	**3**	**2**
CCS	0.477	0.495	0.485	4	6	3
POLY	0.475	0.494	0.484	5	8	5
EXB	0.467	0.495	0.480	9	13	4
NTNU	0.470	0.456	0.462	13	12	17
LCS-IESI	0.461	0.456	0.458	15	15	18
UA-DLSI	0.457	0.456	0.456	17	18	16
Lead	0.425	0.434	0.429	20	24	20

Table 1: MSS task. English.

system	P score	R score	F score	P-Rank	R-Rank	F-Rank
CCS	0.202	0.213	0.207	1	1	1
MUSE	**0.196**	**0.210**	**0.203**	**2**	**2**	**2**
POLY	0.189	0.203	0.196	4	4	6
EXB	0.186	0.205	0.195	5	5	4
Oracles	0.182	0.204	0.192	6	6	5
Lead	0.168	0.178	0.173	12	13	12
LCS-IESI	0.181	0.170	0.172	13	7	14

Table 2: MSS task. Hebrew.

system	P score	R score	F score	P-Rank	R-Rank	F-Rank
Oracles	0.630	0.658	0.644	1	1	1
MUSE	**0.562**	**0.569**	**0.565**	**2**	**4**	**2**
CCS	0.554	0.571	0.562	4	3	3
EXB	0.546	0.571	0.558	8	2	7
POLY	0.545	0.560	0.552	10	9	9
LCS-IESI	0.540	0.527	0.531	11	13	12
Lead	0.524	0.535	0.529	13	12	13

Table 3: MSS task. Arabic

System	R-1 R	R-1 P	R-1 F	R-2 R	R-2 P	R-2 F
Gillick and Favre	0.401	0.407	0.401	0.160	0.162	0.160
WECOM + Gillick and Favre	**0.410***	0.413	0.409	0.166	0.166	0.165
McDonald	0.393	0.407	0.396	0.156	0.159	0.156
WECOM + McDonald	**0.401***	0.403	0.399	0.158	0.158	0.157
POLY + POS_F	0.448	0.453	0.447	0.213	0.214	0.212
WECOM + POS_F	0.450	0.450	0.447	0.211	0.210	0.210

Table 4: ROUGE-1 and -2 scores. DUC2002.