# The participation of UJF-Grenoble team at Multiling 2015

**Georgios Balikas**
University of Grenoble Alpes, France
georgios.balikas@imag.fr

**Massih-Reza Amini**
University of Grenoble Alpes, France
massih-r.amini@imag.fr

## Abstract

This paper describes the UJF-Grenoble's team participation in the Multiling 2015 challenge. Specifically, we participated in the Multilingual, Multi-Document Summarization task for which we implemented an extractive summarization approach. We proposed a method that does not require trained natural language processing resources but big volumes of free text for each language. It consists of a representation learning and a sentence selection step. In the former, it relies on Neural Networks to produce enhanced text representations and in the latter on a cosine similarity model to select the most appropriate sentences to be included in the generated summary.

## 1 Introduction

Over the past years, several challenges and workshops concerning sub-areas of Natural Language Processing (e.g. question answering, named-entity recognition, summarization etc.) have been organized. This year a community-driven initiative organised the Multiling 2015 challenge, which is a set of challenges providing the infrastructure for evaluating multilingual summarization systems in order to push the state-of-the-art in the field. The challenge comprised four different tasks, namely the Multilingual Multi-document Summarization (MMS) task, the Multilingual Single-document Summarization (MSS) task, the Online Forum Summarization (OnForumS) task and the Call Centre Conversation Summarization (CCCS) task. In this paper we present our participation in the Multilingual Multi-document Summarization task. In section 2 we provide a brief description of the MMS task. Section 3 describes in detail our systems while section 4 presents the official results. Finally, section 5 concludes.

## 2 Multilingual, Multi-Document Summarization

The multilingual multi-document summarization task aimed at evaluating the application of language-independent summarization methods on a variety of languages. The participating systems had to generate summaries given a multilingual corpus created by news stories. To demonstrate the applicability of their methods in several languages, participants had to generate summaries in at least two of the following languages: Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian and Spanish.

For each of the above-mentioned languages the test documents was organised in news events. A news event is a set of 10 documents (cleaned, utf-8 encoded text) describing different aspects of a main news story. The original news articles come from http://www.wikinews.org/ and were translated by the organisers in the rest of the languages. The topics of the news events are wide: sports events, human disasters, terrorist attacks etc. Depending on the language 10 or 15 news events were distributed which resulted in 100 or 150 documents and required 10 or 15 summaries to be generated respectively. For enabling the participation of supervised approaches, the organisers also released the golden (human generated) summaries of three of these news events.

## 3 System Overview

We suppose that there exists a set of $K$ news events $\boldsymbol{E} = \{e_i\}_{i=1}^{K}$. A news event $e_i$ is a news story described by $N$ documents, with $e_i =$

$\{\{d_{i,j}^{(\ell)}\}_{\ell=1}^{v}\}_{i=1}^{N}$ where $d_{i,j}^{(\ell)}$ is the *j*-th document of the *i*-th news event in the $\ell$-th language. Respectively, $S_{i,j}^{(\ell)}$ is the set of its sentences and $t_{i,j}^{(\ell)}$ is this document's title. For the case of the Multiling 2015 challenge $N = 10$.

Our approach decomposes in two steps: (i) a representation learning step where the goal is to learn representations of sentences that capture the semantics of a sentence and (ii) a sentence selection step where the most "appropriate" sentences are selected to be included in the summary. We describe below the two steps and we also discuss the concept of "sentence appropriateness" for the summaries.

**Sentence Embeddings.** Let $\mathcal{G}$ be a transformation of a given text span that projects it to a vector space of dimension $d$, where $d$ is user defined (typically between 50-500): for a sentence $s_i \in S_{i,j}^{(\ell)}$, $\mathcal{G}^{(\ell)}(s_i) \subset \mathbb{R}^d$. In the paper we refer to $\mathcal{G}^{(\ell)}(s_i)$ as a distributed representation of $s_i$ or its sentence embedding interchangeably, denoting a dense vector of real-valued features which characterize the meaning of the sentence (Hinton, 1986).

For our Multiling 2015 participation we considered sentence embeddings learnt with neural networks. In (Mikolov et al., 2013) the authors discuss the *continuous bag of words* (cbow) and the continuous *skip-gram* models that generate representations for words. In this case, we define $\mathcal{G}$ by averaging the vector representations of words in a sentence or a query, a process we refer to as `average pooling`.

To generate embeddings of larger text spans without average pooling, (Le and Mikolov, 2014) proposes the *Distributed Memory Model of paragraph vectors* (DMMpv) and the *Distributed Bag-Of-Words of paragraph vectors* (DBOWpv) models, which are extensions of the cbow and the skip-gram respectively. For the latter, $\mathcal{G}$ is defined by the model outputs. It is to be noted that $\mathcal{G}$, dubbed $\mathcal{G}^{(\ell)}$, has to be learnt in advance for each of the languages we are interested in.

**Sentence Extraction** Since we had no previous experience with summarization, we opted for a simple cosine measure as in (Knaus et al., 1995). We decided to select sentences as the text spans to be extracted from the given corpus. Said that, the extractive summarization decouples in (i) ranking the sentences of the documents of an event $e_i$ using the cosine measure and (ii) progressively adding sentences to the summary, starting from the top-ranked. To rank the sentences we compare them against a query $q_i^{(\ell)}$ and the title of the document they come from. The query $q_i^{(\ell)}$ consists of the most frequent terms (excluding stop words) in $e_i$ and aims at capturing the general ideas in $e_i$ in the form of words. The score of a sentence $s_{i,j}^{(\ell)}$ is hence:

$$\text{sc}(s_{i,j}^{(\ell)}) = \alpha\cos(\mathcal{G}(s_{i,j}^{(\ell)}), \mathcal{G}(q_i^{(\ell)})) + (1-\alpha)\cos(\mathcal{G}(s_{i,j}^{(\ell)}), \mathcal{G}(t_{i,j}^{(\ell)})) \quad (1)$$

where $\mathcal{G}(.)$ is the embedding of the sentences and the queries in a common vector space and $\alpha$ is a real-valued mixing hyper-parameter that regulates the contribution of the sentence similarity with the query and the the titles to its score.

We tried two different ways of ranking the sentences of a news event. The first one, dubbed "serial" ranks the sentences of each document separately, thus creating $N$ rankings (one for each of the documents of a news event) for each news event. Then, from each document the higher-scoring sentence is selected which ensures that most of the documents will be represented in the final summary. The second, dubbed "pool", creates a single ranking of the sentences from a news event. This approach relies only on the weighting scheme of Eq. (1) to ensure the maximum coverage of the input documents. The rationale is that some of the sentences of each document will score high in the pool because they will be similar enough both with the title and with the query. In both cases, the sentences are ranked by calculating their scores using Eq. (1).

Two known issues one needs to cope with in MDS are *redundancy* and *discourse incoherence*. The former deals with the fact that the source documents share common information and, therefore, sentences extracted from different source documents may repeat the same information. To deal with this, we added a new sentence $s_{i,j}^{(\ell)}$ to the summary iff:

$$\underset{s_i^{(\ell)} \in \text{Summary}}{\arg\max} \quad \cos(\mathcal{G}(s_{i,j}^{(\ell)}), \mathcal{G}(s_i^{(\ell)})) < \theta \quad (2)$$

where $\theta$ is a hyper-parameter to be tuned. From Eq. (2) a sentence is added in the summary only

**Algorithm 1** "Serial" summarizer

**Require:** $\mathcal{G}^{(\ell)}(s_i)$, $e_i$

1: **for** each document $s_i \in d_{i,j}^{(\ell)}$ **do**
2: Calculate the sentence scores using Eq. (1). Rank them according to those scores.
3: **end for**
4: Order the documents according to their publication timestamp.
5: **for** each document until the summary length is less than 250 words **do**
6: Append the highest-scoring sentence that fulfills Eq. (2) in the summary.
7: **end for**
8: Return the generated summary.

**Algorithm 2** "Pool" summarizer

**Require:** $\mathcal{G}^{(\ell)}(s_i)$, $e_i$

1: **for** each document $s_i \in d_{i,j}^{(\ell)}$ **do**
2: Calculate the sentence scores using Eq. (1).
3: **end for**
4: Rank the sentences of all documents according to their scores.
5: **for** each sentence from the ranked pool until the summary length is less than 250 words **do**
6: Append it in the summary iff it fulfills Eq. (2)
7: **end for**
8: Return the summary after ordering the selected sentences wrt their publication dates.

if its semantic similarity with the already-added sentences is below $\theta$, where $\theta$ is a parameter to be tuned.

Concerning discourse incoherence, in MDS it is unlikely that the extracted sentences will form a coherent and readable text if presented in an arbitrary order. We used a simple heuristic: once we had the set of sentences to be included in the summary, we ordered them using the publication timestamp of the article they came from and we resolved ties using the scores of Eq. (1). Algorithms 1 and 2 sketch the "serial" and the "pool" approach respectively.

## 4 Results and Discussion

Table 1 presents the statistics of the data we used to train the $\mathcal{G}$ transformation, for each of the languages we participated, namely English, French and Greek. The free text for each language comes from news events and is publicly available at `http://statmt.org/wmt12/`. Also, since the dimension of the distributed representations $d$ has to be provided to the algorithms that will generate them, Figure 1 shows a greed search for the dimensions in the set of the 3 summaries that were

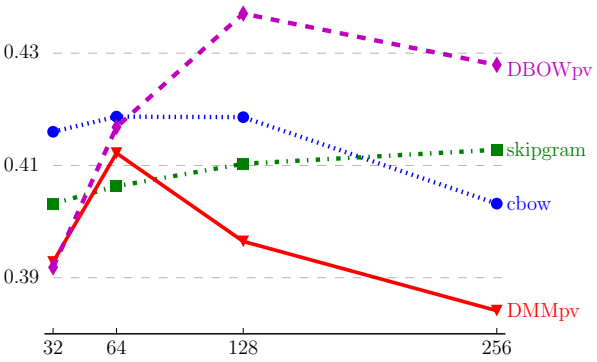


Figure 1: The summarization performance of "pool" for English for different dimensions of the produced distributed representations.

| | Sentences | Vocabulary | # Words |
|---|---|---|---|
| English | 8,139,382 | 899,163 | 255,045,349 |
| French | 5,589,090 | 604,965 | 188,015,178 |
| Greek | 2,320,442 | 258,235 | 59,779,505 |

Table 1: Statistics for the training data we used to generate the word and sentence embeddings.

released by the organisers for the tuning purposes of supervised systems. The scores are presented with regard to the Rouge2 (Lin, 2004) evaluation measure for the English part of the coprus. As a result, we decided to generate representations of dimension 128 using the DBOWpv model. To generate the word and sentence representations of the four models of Figure 1 we used the open implementations of Gensim (Řehůřek and Sojka, 2010).

We submitted three runs in the track. Our first system, `MMS1a` is based on the "serial" algorithm, and the sentece embeddings are generated for each langauge. Our second system `MMS1b` is based on the "pool" algorithm and averages the sentence representations in three languages (English, French and Greek) in an effort to combine information from all three of them. It uses the fact that the multilingual coprus consists of parallel translations of news articles primarily edited in English. Finally, our third system, `MMS1c`, is also based on the "pool" algorithm using sentence embeddings generated for each language.

Tables 2, 3 and 4 present the averages of the summarization scores of the news events of the corpus for English, French and Greek respec-

| System | AutoSummENG | MeMoG | NPowER |
|---|---|---|---|
| MMS8a | 0.1925 | 0.2185 | 1.9046 |
| MMS2 | 0.1909 | 0.2220 | 1.9054 |
| MMS8c | 0.1906 | 0.2159 | 1.8975 |
| MMS8b | 0.1905 | 0.2129 | 1.8937 |
| MMS5a | 0.1778 | 0.1936 | 1.8436 |
| MMS1a | 0.1751 | 0.1988 | 1.8441 |
| MMS15 | 0.1744 | 0.2004 | 1.8446 |
| MMS5b | 0.1708 | 0.1944 | 1.8297 |
| MMS1b | 0.1695 | 0.1960 | 1.8289 |
| MMS11a | 0.1688 | 0.1836 | 1.8125 |
| MMS9 | 0.1657 | 0.1797 | 1.8013 |
| MMS3 | 0.1640 | 0.1848 | 1.8039 |
| MMS1c | 0.1608 | 0.1817 | 1.7933 |
| MMS13a | 0.1607 | 0.1801 | 1.7911 |
| MMS13b | 0.1594 | 0.1780 | 1.7860 |
| MMS11b | 0.1572 | 0.1696 | 1.7712 |
| MMS12 | 0.1475 | 0.1651 | 1.7453 |

Table 2: English: The performance of the participating systems with regard to the three official Multiling 2015 evaluation measures.

tively. The scores are generated for the three official evaluation measures of the challenge: AutoSummENG (Giannakopoulos et al., 2008), MeMoG (Giannakopoulos and Karkaletsis, 2011), and NPowER (Giannakopoulos and Karkaletsis, 2013). The three tables are sorted based on the AutoSummENG measure and our systems are presented in a different font. In the cases of Greek and English language MMS1a performs the best between the three systems. For those languages choosing the best scoring sentence from each input document of a news event yieded good results. Examining the generated summaries manually, we believe that this approach benefited the coverage of the aspects of the substories of each news event. On the other hand, our best performing system for French was MMS1c.

Concerning the comparison of our performance with the rest of the participating systems we are probably ranked in the middle. Our best performing system scored above the average performance for each of the languages we participated. In our case, integrating distributed representations of sentences in an extractive summarization approach yielded encouraging results. With a simple sentence selection method and without using any natural language processing or language-dependent tools we achieved above average performance.

An interesting insight we gained concerns our system MMS1b. From Tables 2 and 3 we notice that it performed between the rest of our systems while using a combination of the representations of the input languages. This averaging of different language representations, in English for instance, improved the performance over the respective monolingual system (MMS1c). We believe that combining representations from different lagnuages to generate more robust ones, can be an interesting research topic.

| System | AutoSummENG | MeMoG | NPowER |
|---|---|---|---|
| MMS2 | 0.2479 | 0.2661 | 2.0792 |
| MMS8b | 0.2229 | 0.2430 | 1.9984 |
| MMS8c | 0.2177 | 0.2241 | 1.9648 |
| MMS8a | 0.2157 | 0.2257 | 1.9624 |
| MMS3 | 0.1987 | 0.1982 | 1.8934 |
| MMS1c | 0.1984 | 0.1863 | 1.8784 |
| MMS9 | 0.1974 | 0.2243 | 1.9220 |
| MMS1b | 0.1924 | 0.2018 | 1.8844 |
| MMS1a | 0.1869 | 0.2108 | 1.8835 |
| MMS5a | 0.1858 | 0.1911 | 1.8575 |
| MMS5b | 0.1826 | 0.1851 | 1.8436 |
| MMS15 | 0.1582 | 0.1743 | 1.7789 |
| MMS12 | 0.1511 | 0.1639 | 1.7514 |

Table 3: French: The performance of the participating systems with regard to the three official Multiling 2015 evaluation measures.

| System | AutoSummENG | MeMoG | NPowER |
|---|---|---|---|
| MMS8c | 0.1623 | 0.1809 | 1.7955 |
| MMS8a | 0.1621 | 0.1823 | 1.7969 |
| MMS9 | 0.1611 | 0.1836 | 1.7962 |
| MMS1a | 0.1575 | 0.1713 | 1.7740 |
| MMS8b | 0.1573 | 0.1700 | 1.7720 |
| MMS2 | 0.1549 | 0.1727 | 1.7701 |
| MMS15 | 0.1497 | 0.1698 | 1.7555 |
| MMS3 | 0.1419 | 0.1625 | 1.7304 |
| MMS12 | 0.1362 | 0.1513 | 1.7048 |
| MMS1c | 0.1357 | 0.1492 | 1.7011 |
| MMS1b | 0.1298 | 0.1337 | 1.6701 |
| MMS5a | 0.1284 | 0.1292 | 1.6618 |
| MMS5b | 0.1267 | 0.1368 | 1.6671 |

Table 4: Greek: The performance of the participating systems with regard to the three official Multiling 2015 evaluation measures.

## 5 Conclusion

We presented our participation in Multiling 2015 challenge in the Multilingual Multi-Document track. We investigated the performance of distributed representations of sentences in an extractive summarization setting. We obtained encour-

aging results which demonstrated that representations learnt using shallow architectures of neural networks can be used to leverage large volumes of free text. In our future work, we aim to find ways (i) to tune the hyper-parameters of our model more efficiently, (ii) to integrate our sentence representations in more complex summarizers and evaluate their performance and (iii) to exploit the multilingual translations to generate representations that will perform better compared to the monolingual approach we introduced here.

## Acknowledgements

This work is partially supported by the CIFRE N 28/2015 and by the LabEx PERSYVAL Lab ANR-11-LABX-0025.

## References

[Giannakopoulos and Karkaletsis2011] George Giannakopoulos and Vangelis Karkaletsis. 2011. Autosummeng and memog in evaluating guided summaries.

[Giannakopoulos and Karkaletsis2013] George Giannakopoulos and Vangelis Karkaletsis. 2013. Summary evaluation: Together we stand npowered. In *Computational Linguistics and Intelligent Text Processing*, pages 436–450. Springer.

[Giannakopoulos et al.2008] George Giannakopoulos, Vangelis Karkaletsis, George Vouros, and Panagiotis Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3):5.

[Hinton1986] Geoffrey E Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA.

[Knaus et al.1995] Daniel Knaus, Elke Mittendorf, Peter Schauble, and Paraic Sheridan. 1995. Highlighting relevant passages for users of the interactive spider retrieval system. In *Proceedings of the fourth text retrieval conference (TREC-4)*, pages 233–244.

[Le and Mikolov2014] Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

[Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Řehůřek and Sojka2010] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.