# Learning language-independent sentence representations for multi-lingual, multi-document summarization

Georgios Balikas*[1,2] and Massih-Reza Amini†[1]

[1]Université Grenoble Alpes, France
[2]Coffreo, Clermont Ferrand

## Abstract

This paper presents an extension of a denoising auto-encoder to learn language-independent representations of parallel multilingual sentences. Each sentence from one language is represented using language dependent distributed representations. The input of the auto-encoder is then constituted of a concatenation of the distributed representations corresponding to the vector representations of translations of the same sentence in different languages. We show the effectiveness of the learnt representation for extractive multi-document summarization, using a simple cosine measure that estimates the similarity between vectors of sentences found by the auto-encoder and the vector representation of a generic query represented in the same learnt space. The top ranked sentences are then selected to generate the summary. Compared to other classical sentence representations, we demonstrate the effectiveness of our approach on the TAC 2011 MultiLing collection and show that learning language-independent representations of sentences that are translations one from another helps to significantly improve performance with respect to Rouge-SU4 measure.

## 1 Introduction

With the explosion of on-line text resources, it has become necessary to provide users with systems that obtain answers to queries efficiently and effectively. In various information retrieval (IR) tasks, multi-document text summarisation (MDS) systems are designed to help users to quickly find the needed information. For example, MDS can be coupled with conventional search engines and help users to evaluate the relevance of groups of documents for providing answers to their queries.

Automated summarization dates back to the fifties [21]. The different attempts in this field have shown that human-quality text summarization was very complex since it encompasses discourse understanding, abstraction, and language generation [14]. Simpler approaches were then explored which consist in extracting representative text-spans, using statistical techniques and/or techniques based on surface domain-independent linguistic analyses. Within this context, summarization can be defined as the selection of a subset of the document sentences which is representative of its content. This is typically done by ranking document sentences and selecting those with higher score and minimum overlap [4]. Most of the recent work in summarization uses this paradigm. Usually, sentences are used as text-span units but paragraphs have also been considered [23]. The latter may sometimes appear more appealing since they contain more contextual information. The quality of an extract summary might not be as good as an abstract summary, but it is considered good enough for a reader to understand the main ideas of a set of documents.

Additionally, people are now confronted with event news available in more than one language. This is a typical situation in many multilingual regions of the world, including many regions of Europe and, for example, the Syndicate project[1] provides commentaries of news events translated in several languages. However, MDS models are

*georgios.balikas@imag.fr, gbalikas@coffreo.com
†massih-reza.amini@imag.fr

[1]http://www.project-syndicate.org/

mostly developed in a monolingual context, typically for English documents.

The situation we are investigating in this paper is when news events are available in more than one language and are translations of one from another. In that case, it is obviously possible to design monolingual MDS for each language independently. The challenge is actually to come up with a method which is able to leverage the multilingual data in order to produce a better performing system than what one gets from the independent monolingual MDS systems alone.

To tackle this problem, our work takes the text-span extraction paradigm and explores a machine learning approach for improving the vectorial representation of multilingual text-spans, that we consider here as being sentences. We propose to learn the representation of multilingual sentences using a denoising auto-encoder (dAE). The input of the auto-encoder is constituted of a concatenation of original input vectors of sentences corresponding to the vector representations of translations of the same sentence in different languages. We use the dAE to learn a language independent vector of the concatenated input vector and in order to force the hidden layer to discover more robust features and prevent it from simply learning the identity, we train the auto-encoder to reconstruct the input from a corrupted version of it. In order to show the impact of sentence representation in the performance of a given similarity based statistical model, we consider different scenarios of mono and multi-lingual sentence representations proposed in the literature. We show that compared to the existing solutions for sentence characterisation, the language-independent sentence representation learned by DAE leads to a significant increase in performance of the statistical model.

In the rest of the paper, after presenting the related work in Section 2, we introduce the proposed representation learning of multilingual sentences as well as different scenarios for multilingual MDS in Section 3. Then, we present our evaluation framework in Section 4 and we conclude and discuss extensions of this work in Section 5.

## 2 Related Work

The original problem of summarization requires the ability to understand and synthesise a document in order to generate its abstract. However, different attempts to produce human quality summaries have shown that this process of abstraction is highly complex, since it needs to borrow elements from fields such as linguistics, discourse understanding and language generation [26]. Instead, most studies consider the task of text summarization as the extraction of text spans (typically sentences) from the original document; scores are assigned to text units and the best-scoring spans are presented in the summary. These approaches transform the problem of summarization into a simpler problem of *ranking* spans from an original text according to their relevance to be part of the document summary. This kind of summarization is related to the task of document retrieval, where the goal is to rank documents from a text collection with respect to a given query in order to retrieve the best matches. Although such an extractive approach does not perform an in-depth analysis of the source text, it can produce summaries that have proven to be effective [17, 29].

To compute sentence scores, most previous studies adopt a linear weighting model which combines simple statistical or linguistic features characterising each sentence in a text [22]. In many systems, the set of feature weights are tuned manually; this may not be tractable in practice, as the importance of different features can vary for different text genres [9]. Machine Learning (ML) approaches within the classification framework, have shown to be a promising way to combine automatically sentence features [16, 30, 5, 2]. In such approaches, a classifier is trained to distinguish between two classes of sentences: summary and non-summary ones. The classifier is learnt by comparing its output to a desired output reflecting global class information. This framework is limited in that it makes the assumption that all sentences from different documents are comparable with respect to this class information.

Word embeddings, introduced by [3] are parametrized functions for mapping words in high (typically 100-500) dimension space. They build on the idea of distributed representations intro-

duced by Hinton [10] where one relies on a neural network to discover features that characterize the meaning of a concept. Recent work has shown that using large amounts of text one can generate such distributed representations of words [24, 27], phrases [25] and even paragraphs or documents [19] in an unsupervised way. In those studies the embeddings were shown to capture the semantics of the text volumes they modelled; using them the authors also improved the state-of-the-art in several ML and IR tasks. Here we explore a multi-view approach for text summarization by adapting the embedding approaches for sentence representation to the multilingual case.

Lastly, our approach bears some similarities with the study described in [15] aiming at the evaluation of continuous vectors space models in extractive summarization. The main difference between this study and ours lies in the fact that our goal is to improve summarization performance using a multi-view approach by exploiting text written in several languages whereas [15] focuses on single-language scenarios. Also, some recent, innovative studies on Machine Translation (MT) [7, 18] adopt a similar idea of projecting phrases or words in a language independent space. Although we project the sentences in a language-independent space our goal is different; we aim to discover a space to efficiently extract sentences for our summaries whereas the MT approaches aim to find more accurate translations.

# 3 Multilingual MDS model

In this section we present the summarization model we consider in order to evaluate the impact of the representation learning strategy we propose. We begin by describing the framework and the summarization model, and then we present three standard sentence representations as well as our approach.

## 3.1 Framework and MDS Model

Here we suppose that there exists a set of $K$ news events $E = \{e_i\}_{i=1}^{K}$. A news event $e_i$ is described by a set of $N$ documents, $D_i = \{\boldsymbol{d_{i,j}}\}_{j=1}^{N}$. Hence, $\boldsymbol{d_{i,j}}$ is the *j-th* document describing the *i-th* news event. Each document $\boldsymbol{d_{i,j}}$ is described by a set of $v$ independent views, such that $\boldsymbol{d_{i,j}} = \{d_{i,j}^{(k)}\}_{k=1}^{v}$ where each view is the document in a different language. For example, if the documents are available in two languages, English and French, then $\boldsymbol{d_{i,j}} = \{d_{i,j}^{(1)}, d_{i,j}^{(2)}\}$, where $d_{i,j}^{(1)}$ is the English version and $d_{i,j}^{(2)}$ is the French version of the same document. We denote $S_{i,j}^{(v)}$ the set of sentences of the document $d_{i,j}^{(v)}$. We finally assume that the sentences between the views of a language are aligned between them i.e., $|S_{i,j}^{(1)}| = |S_{i,j}^{(2)}|$ and the first sentence of $S_{i,j}^{(1)}$ has the same meaning with the first sentence of $S_{i,j}^{(2)}$ etc., which is common when translating documents from one language to another. The operator $|.|$ returns the cardinality of a set. Note that we do not assume that the words in the sentences are aligned as in machine translation problems [31].

Many systems for sentence extraction rely on the use of similarity measures between text spans (sentences in our case) and queries, e.g. [12, 13]. The extractive summarization then decouples in (i) ranking the sentences of the documents $D_i$ using a scoring function and (ii) progressively selecting sentences to be added to the summary, starting from the top-ranked.

These systems differ in the representation of textual information and in the similarity measures they use. Usually, statistical and/or linguistic characteristics are used in order to encode the text (sentences and queries) into a fixed size vector and simple similarities are then computed.

**Similarity measure.** In order to avoid the bias of the similarity model in our analysis, we build on the work of [6] who used a simple cosine measure for the extraction of sentences relevant to a generic query. In our experiments we considered a generic query constituted of the most frequent terms in a document $D_i$ that are not no-stop words, denoted by $q_i$, and its title $t_i$. The similarity of a sentence $s \in D_i$ and the generic query is then defined as

$$\text{score}(s) = \alpha \cos(\mathcal{G}(s), \mathcal{G}(q_i)) + (1-\alpha)\cos(\mathcal{G}(s_i), \mathcal{G}(t_i)) \quad (1)$$

where, $\mathcal{G}(.)$ is a vector representation of sentences and queries in the same vectorial space (described in Section 3.2) and $\alpha$ is a real valued mixing hyper-parameter.

**Redundancy.** In the case of MDS, the source documents share common information and, therefore, sentences extracted from different source documents may repeat the same information. To overcome this, we build on the work of [13] and require that sentences to be added in a summary should have small content overlap with previously chosen sentences. Formally, we add a new sentence $s_i$ to the summary that contains the sentence $s_j$ iff:

$$\arg\max_{s_i \in \text{Summary}} \cos(\mathcal{G}(s), \mathcal{G}(s_i)) < \theta \quad (2)$$

where $\theta$ is a hyper-parameter to be tuned. As a result, having the ranked sentences, we begin with the top-ranked and we add sentences to the summary as long as they fulfil the redundancy criteria.

**Discourse incoherence.** In the case of MDS it is unlikely that the extracted sentences will form a coherent and readable text if presented in an arbitrary order. We tackle here this problem with a simple heuristic approach: Once the initial set of sentences is found from the previous step, we re-rank the sentences that will constitute the summary using the date of the article they come from and resolve ties using the scores of Equation (1).

### 3.2 Sentence embeddings

In this section, we present the embeddings we consider, and propose, for implementing the transformation, $\mathcal{G}$, that projects a sentence or a query into a vector space of dimension $d$, where $d$ is user defined, *i.e.,* given a sentence $s_i \in S_{i,j}^{(v)}$, $\mathcal{G}^{(v)}(s_i) \subset \mathbb{R}^d$.

**GLObal VEctors** (Glove) proposed by [27] consider the co-occurrences of words in a large corpus and make the assumption that words occurring in the same documents with the same frequencies are similar and should lie close in the embedding's vector space. This hypothesis was successfully applied to find word representations for text summarization in [1]. Following this idea, [27] used a regression model to find word vectors. In this case, we define the transformation $\mathcal{G}$ by averaging the vector representations of words in a sentence or a query, a process that we refer to as `average pooling`:

$$\mathcal{G}(x) = \frac{1}{|x|} \sum_{w \in x} \mathbf{w}$$

where, $x$ represents either a query or a sentence, $w$ is a word within $x$ and $\mathbf{w}$ is its corresponding embedding.

**Continuous Bag Of Words** (cbow) [24] also learns word embeddings using this time a neural network model, that is built to predict a word within its context defined as words before and after it. The word embeddings are initialized randomly but as training proceeds they eventually capture the semantics of the words as an indirect result of the word prediction task. Once the word embeddings are learnt, the transformation $\mathcal{G}$ is defined by average pooling as in the previous case.

**Distributed Memory Model of paragraph vectors** (DMMpv) is an extension of cbow proposed by [19] to directly learn embeddings for larger text spans than words, such as phrases, sentences or paragraphs. Apart from the word embeddings, the input of the Neural Networks includes a token for the text-span to which the words belong to. This token can be seen as another word, but it actually acts as an identity (or memory) of the content of the text-span. Note that the word representations are shared across the training corpus whereas each text-span has its own token. In this case, the transformation $\mathcal{G}$ is defined over the output of the model.

**Multilingual Language Independent Embedding** (MLIE) model that we propose is an extension of the single-language sentence embedding to the multilingual case. The model is based on the hypothesis that different translations of a sentence employ synonym words across languages and hence bring more information on its content than each translation alone. To achieve that, we project multilingual representations of sentences into a language-independent space using an autoencoder (AE). Figure 1, illustrates this for three languages, English, French and Greek.

AEs [11] are a family of feed-forward neural networks that are trained to reconstruct the input data by performing two steps. In the first step, an input vector from $\mathbb{R}^d$ is projected to a space $\mathbb{R}^a$,

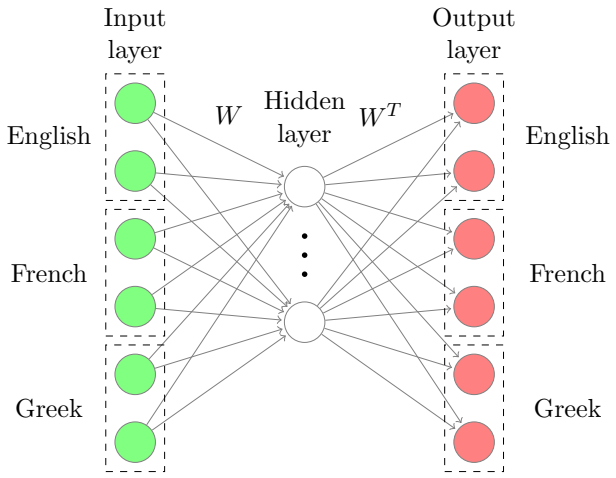


Figure 1: Mutlilingual sentence embedding with an autoencoder that projects its inputs to a language independent space (hidden layer encoding). The inputs here are the concatenated representations of a sentence in three different languages: English, French and Greek. In the figure, each dashed box represents the vector representation of dimension $d$ of a sentence in the corresponding language. A stochastic corruption noise is applied on the input vectors.

called encoding, using non-linear, bijective functions, where usually $a < d$. In the second step, the encoded vector is projected into the original space of dimension $d$ using again non-linear, bijective functions. The AE model that we developed is trained using the stochastic back-propagation algorithm in order to minimize the reconstruction error between the input and output predicted vectors.[2] After training, all the necessary information of the input vector is contained in the compressed representation of the encoding.

In order to force the hidden layer to discover more robust features and prevent it from simply learning the identity, we also train the AE to reconstruct the input from a corrupted version of it. The network then tries to learn an encoding of the input while a corruption process is stochastically applied to it, hence called denoising auto-encoder (dAE). The applied corruption can also be seen as a form of regularization that prevents over-fitting the training data. Formally, if $\boldsymbol{x}$ is the input vector to the dAE, the encoded vector has the form $\boldsymbol{y} = s(\boldsymbol{W}\hat{\boldsymbol{x}} + \boldsymbol{b})$, where $\hat{\boldsymbol{x}}$ is the corrupted input, $\boldsymbol{W}$ are the weights that link the input nodes to the hidden-layer nodes, $\boldsymbol{b}$ is the bias, and $s$ is the activation function commonly taken as the logistic or the hyperbolic tangent function. Similarly, for the decoding step we have $\boldsymbol{z} = s(\boldsymbol{W'y} + \boldsymbol{b'})$. The network is trained to learn the weights of the links $\boldsymbol{W}$ and $\boldsymbol{W'}$ and the biases $\boldsymbol{b}$ and $\boldsymbol{b'}$, by minimizing the average of the euclidean distance between the predicted vectors, $\boldsymbol{z}$, and the input vectors, $\boldsymbol{x}$, of a training set.

For learning the sentence embeddings, we consider the input vector as the concatenation of the vector representations of each sentence among different languages. For each multilingual sentence and generic query we, hence, use their language independent vector representations obtained after encoding to generate the summaries across languages as for the monolingual case. It is to be noted that there is no restriction on the number of the input languages; in the described setting however, the bigger the number of the input languages the more the free-parameters to be learned. The total number is $2 \times a \times d \times v$, where $v$ is the number of the input languages, $d$ is the dimension of the sentence embeddings in each language and $a$ is the number of the units in the hidden layer. In our approach we use tied weights i.e., $W = W^T$, to reduce the free parameters by a fraction of 2.

## 4 The Experimental Framework

**The data.** We begin by describing the details of our experimental setup and then present our results. The data we used come from a pilot study for multi-document, multi-lingual summarization [8]. In the framework of the study, a dataset was created by gathering an English corpus of 10 topics and then translating it into 6 other languages using a sentence-by-sentence approach. Each topic is the basis of a news event that contains 10 relevant articles. In total, the summarization problem for each language has 100 documents: 10 documents for each of the 10 news events. We use here the fraction of the dataset that corresponds to English, French and Greek.

Table 1 presents the details for the data we used to generate the word and the phrase embeddings. They consist of sentences from news arti-

[2]We will make the code publicly available.

| | Sentences | Vocabulary | Size |
|---|---|---|---|
| English | 8,139,382 | 899,163 | 1.3 Gb |
| French | 5,589,090 | 604,965 | 1.2 Gb |
| Greek | 2,320,442 | 258,235 | 696 Mb |

Table 1: Statistics for the training data we used to generate the word and the sentence embeddings. We report the numbers of the sentences in the training set, the vocabulary size and the size of the training files (uncompressed).

cles.[3] To generate the vectors produced by the DMMpv model we used the Gensim implementation [28], with distributed memory, hierarchical sampling and window size of 8 words and $d = 128$. To generate the Glove and the cbow representations we used the publicly available implementations provided by the authors of the corresponding papers; concerning the hyper-parameters of the models we used the defaults ones and we generated word embeddings for $d = 50$. The considered dimensions for both DMMpv and cbow are those that provided the best results for these representations. For our AE model, we used hyperbolic tangent activation functions and set the hidden layer neurons to 60% of the input neurons. We used a fraction of the dataset for which the organizers had released the corresponding gold summaries to tune the threshold hyper-parameter $\theta$ (Eq. 2). Finally, the length of the query associated to event $e_i$ was set to the average length of the sentences in the associated documents $D_i$.

To foster evaluation, the organisers also released human-generated, golden summaries for each of the news events after the competition as well as the submissions of the systems that participated in the pilot study. We present the scores of those submissions in Table 2. In this study, we report scores for ROUGE-SU4 [20], which is a recall based measure used in the pilot study and also commonly employed for evaluating summarization systems.[4]

We compared the results of the summarization system described in Section 3.1 using the monolingual sentence embeddings (Glove, cbow and DMMpv) as well as the proposed multilingual sentence representation (MLIE) with each of the latter used in its input by considering the classical case and the denoising case (denoted by $'$ in its exponent). In order to evaluate the benefits of learning the language independent representation using the AE, we also made experiments by concatenating the monolingual representations (the input of AE) denoted by (CONCAT). Table 3 presents the average of ROUGE-SU4 for each case and for different values of $\alpha \in \{-1, -.75, \ldots, 1.75, 2\}$ (Eq. 1). In the parenthesis, we also report the maximum performance across the different values of $\alpha$.

| Rank | ID | English | French | Greek |
|---|---|---|---|---|
| 1 | ID 4 | **0.454** | 0.391 | 0.713 |
| 2 | ID 2 | 0.438 | **0.405** | 0.707 |
| 3 | ID 9 | 0.428 | 0.375 | **0.721** |
| 4 | ID 5 | 0.404 | 0.314 | - |
| 5 | ID 1 | 0.400 | 0.374 | 0.680 |
| 6 | ID 7 | 0.384 | 0.378 | 0.668 |
| 7 | ID 8 | 0.391 | - | - |
| 8 | ID 3 | 0.373 | 0.381 | 0.637 |
| 9 | ID 6 | 0.284 | 0.255 | - |
| 10 | ID 10 | 0.249 | 0.289 | 0.404 |

Table 2: The ROUGE-SU4 scores of the participating systems in the TAC 2011 Multiling pilot study. The systems are ranked according to their performance on the English language. The best performance per language is dubbed bold. The performance of systems that did not submit results for one of the languages is replaced by a dash.

Considering the first three lines of the table corresponding to the mono-lingual case, it comes that the simple summarization strategy relying on the cosine similarity measure without any linguistic knowledge that we propose is competitive with more complex systems relying on such knowledge and that participated to the competition (Table 2). This shows the important role of learning sentence representation for this task. Further, we found that the cbow representation is the best performing compared to the two other monolingual sentence representation. Examining the scores, the performance for the Greek language seems to be far more higher. We believe that this is ought to the internals of the ROUGE package; we have not integrated to it language-dependent tokeniz-

[3]The dataset of the news articles is available at `http://www.statmt.org/wmt10/translation-task.html`.

[4]We report ROUGE scores for summaries of 250 words after running `ROUGE-1.5.5.pl -a -x -2 4 -u -c 95 -e <ROUGEDIR> -r 1000 -n 2 -f A -p 0.5 -t 0 -d <SETTINGS.XML>`

| | | English | French | Greek |
|---|---|---|---|---|
| Mono-L. | cbow | .400 ± .007 (.410) | .372 ± .016 (.395) | .655 ± .005 (.663) |
| | Glove | .390 ± .016 (.413) | .357 ± .007 (.372) | .644 ± .009 (.659) |
| | DMMpv | .387 ± .005 (.393) | .345 ± .002 (.349) | .651 ± .010 (.670) |
| 2 lang. input | $\text{CONCAT}_{\text{cbow}}$ | .396 ± .007 (.405) | .378 ± .010 (.391) | - |
| | $\text{CONCAT}_{\text{Glove}}$ | .382 ± .010 (.404) | .362 ± .013 (.386) | - |
| | $\text{CONCAT}_{\text{DMMpv}}$ | .391 ± .005 (.402) | .350 ± .003 (.354) | - |
| | $\text{CONCAT}_{\text{cbow}}$ | .400 ± .008 (.412) | - | .652 ± .008 (.663) |
| | $\text{CONCAT}_{\text{Glove}}$ | .383 ± .008 (.393) | - | .640 ± .011 (.662) |
| | $\text{CONCAT}_{\text{DMMpv}}$ | .397 ± .008 (.412) | - | .646 ± .007 (.655) |
| | $\text{CONCAT}_{\text{cbow}}$ | - | .376 ± .009 (.391) | .643 ± .006 (.654) |
| | $\text{CONCAT}_{\text{Glove}}$ | - | .356 ± .005 (.363) | .643 ± .012 (.659) |
| | $\text{CONCAT}_{\text{DMMpv}}$ | - | .354 ± .004 (.363) | .647 ± .010 (.658) |
| | $\text{MLIE}'_{\text{cbow}}$ | .401 ± .011 (.414) | .380 ± .010 (.400) | - |
| | $\text{MLIE}_{\text{cbow}}$ | .399 ± .007 (.410) | .380 ± .007 (.390) | - |
| | $\text{MLIE}'_{\text{cbow}}$ | .409 ± .007 (.421) | - | .658 ± .009 (.665) |
| | $\text{MLIE}_{\text{cbow}}$ | .402 ± .009 (.412) | - | .654 ± .007 (.667) |
| | $\text{MLIE}'_{\text{cbow}}$ | - | .369 ± .021 (.390) | .643± .014 (.668) |
| | $\text{MLIE}_{\text{cbow}}$ | - | .371 ± .011 (.383) | .646 ± .011 (.667) |
| | $\text{MLIE}'_{\text{Glove}}$ | .398 ± .021 (<u>.431</u>) | .362 ± .016 (.391) | - |
| | $\text{MLIE}_{\text{Glove}}$ | .377 ± .014 (.401) | .355 ± .015 (.386) | - |
| | $\text{MLIE}'_{\text{Glove}}$ | .404 ± .017 (.427) | - | .665 ± .010 (.681) |
| | $\text{MLIE}_{\text{Glove}}$ | .385 ± .020 (.418) | - | .650 ± .021 (.681) |
| | $\text{MLIE}'_{\text{Glove}}$ | - | .357 ± .015 (.380) | .659 ± .012 (.672) |
| | $\text{MLIE}_{\text{Glove}}$ | - | .349 ± .009 (.367) | .635 ± .007 (.650) |
| | $\text{MLIE}'_{\text{DMMpv}}$ | .413 ± .009 (.426) | .363 ± .010 (.378) | - |
| | $\text{MLIE}_{\text{DMMpv}}$ | **.417** ± .011 (.429) | .367 ± .010 (.378) | - |
| | $\text{MLIE}'_{\text{DMMpv}}$ | .389 ± .008 (.397) | - | .654 ± .008 (.670) |
| | $\text{MLIE}_{\text{DMMpv}}$ | .389 ± .009 (.403) | - | .655 ± .010 (.669) |
| | $\text{MLIE}'_{\text{DMMpv}}$ | - | .354 ± .005 (.363) | .665 ± .010 (<u>.685</u>) |
| | $\text{MLIE}_{\text{DMMpv}}$ | - | .355 ± .005 (.363) | **.666** ± .010 (.683) |
| 3 lang. input | $\text{CONCAT}_{\text{Glove}}$ | .383 ± .009 (.393) | .368 ± .013 (.382) | .641 ± .013 (.660) |
| | $\text{CONCAT}_{\text{DMMpv}}$ | .380 ± .004 (.387) | .344 ± .008 (.353) | .648 ± .013 (.674) |
| | $\text{CONCAT}_{\text{cbow}}$ | .386 ± .006 (.394) | .369 ± .008 (.388) | .648 ± .003 (.653) |
| | $\text{MLIE}'_{\text{cbow}}$ | .410 ± .010 (.422) | .382 ± .016 (.400) | .651 ± .011 (.665) |
| | $\text{MLIE}_{\text{cbow}}$ | .403 ± .009 (.419) | **.384** ± .017 (<u>.409</u>) | .650 ± .009 (.661) |
| | $\text{MLIE}'_{\text{Glove}}$ | .397 ± .008 (.411) | .356 ± .009 (.367) | .664 ± .008 (.676) |
| | $\text{MLIE}_{\text{Glove}}$ | .372 ± .011 (.394) | .346 ± .010 (.365) | .647 ± .008 (.659) |
| | $\text{MLIE}'_{\text{DMMpv}}$ | .388 ± .007 (.398) | .348 ± .006 (.359) | .651 ± .007 (.663) |
| | $\text{MLIE}_{\text{DMMpv}}$ | .384 ± .006 (.393) | .349 ± .005 (.358) | .655 ± .008 (.667) |

Table 3: The performance of the different models on ROUGE-SU-4. In the first (top) part of the table we present the summarization performance when only one language is available in the input. In the middle and the bottom part two and three languages are available respectively. We dub bold the best average performance across the values of $\alpha$ per language and underline the maximum performance obtained in our experiments.

ers/stemmers etc. However, the behaviour across the languages is consistent with respect to the different models and the improvements we obtained.

For the multilingual case, we considered bilingual (2 languages out of 3) and tri-lingual cases. We first notice that learning the latent representation for the multilingual case, improves the summarization performance compared to the single-language approach. Also, the languages with the less training resources, Greek, benefit more compared to their single-language summarization performance. For instance, in the bilingual experiments with $\text{MLIE}'_{\text{cbow}}$ for the language pairs that involved English, the French and the Greek summaries gained 1.3 and 2.3 ROUGE points respectively while the English gained only 0.1. The results are especially interesting as the simple concatenation of monolingual representation leads to a decrease of the performance of the summarization system compared to the initial mono-lingual setting. We believe the AE achieves to disentangle the language-dependent factors of the sentences and captures the semantics of the sentences using the non-linear activation function; this in turn leads to the performance improvements. Finally, it is to be noted that although paragraph vectors did not perform as good in the monolingual setting, in the bilingual setting the performance of the model based on this representation is improved a lot. They are the best performing method (wrt the average scores) for the English and the Greek language throughout our experiments. For French, counter-intuitively, the best performance was obtained with the three languages in the input.

Comparing the best performance we obtained per language (depicted underlined in Table 3) with the performance of the systems that participated to the Multiling 2011 pilot study, we notice that the multilingual learning strategy, we propose, allows the simple cosine based summarization system to end up in the top-3 systems of the challenge. Hence, by employing a simple summarization approach and taking advantage of the enhanced sentence representations we are able to perform well without resorting to complex linguistics or heuristics and without adapting the system to the specificities of one language.

# 5 Conclusion

We proposed a multi-view AE model for learning language independent representation for multilingual sentences. We demonstrated the effectiveness of the proposed approach for the multi-lingual MDS task and showed that it allows to improve the performance of a simple similarity based summarization system comparatively to other complex summarization systems.

# References


[1] Massih Amini and Nicolas Usunier. A contextual query expansion approach by term clustering for robust text summarization. In *Proceedings of the 7th Document Understanding Conference*, Rocherster - USA, 2007.

[2] Massih-Reza Amini and Patrick Gallinari. The use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of the* $25^{th}$ *ACM SIGIR Conference*, pages 105–112, 2002.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[4] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, 1998.

[5] Wesley T. Chuang and Jihoon Yang. Extracting sentence segments for text summarization: a machine learning approach. In *Proceedings of the* $23^{rd}$ *ACM SIGIR Conference*, pages 152–159, 2000.

[6] Knaus D., Mittendorf E., Schauble P., and Sheridan P. Highlighting relevant passages for users of the interactive spider retrieval system. In *TREC-4 proceedings*, 1994.

[7] Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. Learning continuous phrase representations for translation modeling. *Proc.*

*of ACL. Association for Computational Linguistics, June*, 2014.

[8] George Giannakopoulos, Mahmoud El-Haj, Benoît Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. Tac 2011 multiling pilot overview. 2011.

[9] Udo Hahn and Inderjeet Mani. The challenges of automatic summarization. In *IEEE Computer Society*, volume 33, pages 29–36, 2000.

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

[11] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[12] Mani I. and Bloedorn E. Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National Conference on AI*, pages 821–826, 1998.

[13] Goldstein J., Kantrowitz M., Mittal V., and Carbonell J. 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*, SIGIR '99, pages 121–127, 1999.

[14] Karen Sparck Jones. Discourse modeling for automatic summarizing. Technical Report 29D, Computer laboratory, university of Cambridge, 1993.

[15] Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL*, pages 31–39, 2014.

[16] Julian Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In *Proceedings of the* $18^{th}$ *ACM SIGIR Conference*, pages 68–73, 1995.

[17] Adenike M. Lam-Adesina and Gareth J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the* $24^{th}$ *ACM SIGIR Conference*, pages 1–9, 2001.

[18] Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

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

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

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

[22] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal*, 2:159–165, 1958.

[23] Mitra M., Singhal A., and Buckley C. Automatic text summarization by paragraph extraction. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 31–36, 1997.

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

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[26] Chris Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186, 1990.

[27] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12, 2014.

[28] Radim Řehůřek and Petr Sojka. 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 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[29] Tetsuya Sakai and Karen Sparck Jones. Generic summaries for indexing in information retrieval. In *Proceedings of the* $24^{th}$ *ACM SIGIR Conference*, pages 190–198, 2001.

[30] Simone Teufel and Marc Moens. Sentence extraction as a classification task. In *Proceedings of the Intelligent Scalable Text Summarization Workshop, ACL*, pages 58–65, 1997.

[31] Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013.