Centroid-based Text Summarization through Compositionality of Word Embeddings

Gaetano Rossiello **Pierpaolo Basile** **Giovanni Semeraro**
Department of Computer Science
University of Bari, 70125 Bari, Italy
{firstname.secondname}@uniba.it

Abstract

The textual similarity is a crucial aspect for many extractive text summarization methods. A bag-of-words representation does not allow to grasp the semantic relationships between concepts when comparing strongly related sentences with no words in common. To overcome this issue, in this paper we propose a centroid-based method for text summarization that exploits the compositional capabilities of word embeddings. The evaluations on multi-document and multilingual datasets prove the effectiveness of the continuous vector representation of words compared to the bag-of-words model. Despite its simplicity, our method achieves good performance even in comparison to more complex deep learning models. Our method is unsupervised and it can be adopted in other summarization tasks.

1 Introduction

The goal of text summarization is to produce a shorter version of a source text by preserving the meaning and the key contents of the original text. This is a very complex problem since it requires to emulate the cognitive capacity of human beings to generate summaries. Thus, text summarization poses open challenges in both natural language understanding and generation. Due to the difficulty of this task, research work in the literature focused on the *extractive* aspect of summarization, where the generated summary is a selection of relevant sentences from a document (or a set of documents) in a copy-paste fashion. A good extractive summarization method must satisfy and optimize both coverage and diversity properties, where the selected sentences should cover a sufficient amount of topics from the original source text, avoiding the redundancy of information in the summary. The diversity property is fundamental especially for a multi-document summarization. For instance in a news aggregator, a selection of too similar sentences may compromise the quality of the generated summary.

An extractive method should define a sentence representation model, a technique for assigning a score to each sentence in the original source and a ranking module to properly select the most relevant sentences by relying on a similarity function. Following this vision, several summarization methods proposed in the literature use the bag of words (BOW) as representation model for the sentence scoring and selection modules (Radev et al., 2004; Erkan and Radev, 2004; Lin and Bilmes, 2011). Despite their proven effectiveness, these methods rely heavily on the notion of similarity between sentences, and a BOW representation is often not suitable to grasp the semantic relationships between concepts when comparing sentences. For example, taking into account the following two sentences *"Syd leaves Pink Floyd"* and *"Barrett abandons the band"*, in the BOW model their vector (sparse) representations result orthogonal since they have no words in common, nonetheless the two sentences are strongly related.

In attempt to solve this issue, in this work we propose a novel and simple extractive summarization method based on the geometric meaning of the centroid vector of a (multi) document by taking advantage of compositional properties of the word embeddings (Mikolov et al., 2013b). Empirically, we prove the effectiveness of word embeddings with a fair comparison to the BOW representation by limiting, as much as possible, the parameters and the complexity of the method. Surprisingly, the results achieved from our method on the gold standard DUC-2004 dataset are comparable,

Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, pages 12–21,
Valencia, Spain, April 3, 2017. ©2017 Association for Computational Linguistics

and in some cases better, to those obtained using a more complex sentence representations coming from the deep learning models.

In the following section we provide a brief description of word embeddings and text summarization methods. The centroid-based summarization method that uses word embeddings is described in Section 3, followed by experimental results in Section 4. Final remarks and a discussion about our future plans are reported in Section 5.

2 Related Work

2.1 Word Embeddings

Word embedding stands for a continuous vector representation able to capture syntactic and semantic information of a word. Several methods have been proposed in order to create word embeddings that follow the Distributional Hypothesis (Harris, 1954). In our work we use two models[1], continuous bag-of-words and skip-gram, introduced by (Mikolov et al., 2013a). These models learn a vector representation for each word using a neural network language model and can be trained efficiently on billions of words. Word2vec allows to learn complex semantic relationships using simple vectorial operators, such as vec(*king*) $-$ vec(*man*) $+$ vec(*woman*) $\approx$ vec(*queen*) and vec(*Barrett*) $-$ vec(*singer*) $+$ vec(*guitarist*) $\approx$ vec(*Gilmour*). However, our method is general and other approaches for building word embeddings can be used (Goldberg, 2015).

2.2 Text Summarization

Since the first method proposed by (Luhn, 1958), automatic text summarization has been widely addressed by the research community with the proposal of different methodologies as well as toolkits (Saggion and Gaizauskas, 2004). Good surveys are proposed by (Jones, 2007; Saggion and Poibeau, 2013). Since our method exploits word embeddings as alternative representation to BOW, here we focus on the methods sharing this feature. Methods based on matrix factorization, such as Latent Semantic Analysis (LSA) (Ozsoy et al., 2011) and Non-Negative Matrix Factorization (NMF) (Lee et al., 2009), have the aim to arise the latent factors by producing dense and compact representations of sentences. Recently, riding the wave of prominent results of modern Deep Learning (DL) models in many natural language processing tasks (LeCun et al., 2015; Goodfellow et al., 2016), several groups have started to exploit deep neural networks for both *abstractive* (Rush et al., 2015; Nallapati et al., 2016) and *extractive* (Kågebäck et al., 2014; Cao et al., 2015; Cheng and Lapata, 2016) text summarization.

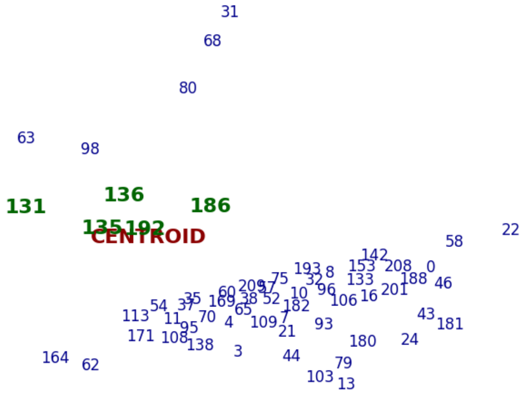

Figure 1: Sentence and centroid embeddings 2D visualization of the *Donkey Kong (video game)* Wikipedia article. Dimensionality reduction is performed using t-SNE algorithm. For each sentence the position in the document is shown. The closest sentences to centroid embedding are marked in green.

3 Centroid-based Method

The centroid-based method for extractive summarization was introduced by (Radev et al., 2004). The centroid represents a pseudo-document which condenses the meaningful information of a document[2]. The main idea is to project in the vector space the vector representations of both the centroid and each sentence of a document. Then, the sentences closer to the centroid are selected. The original method adopts the BOW model for the vector representations using the $tf * idf$ weight scheme (Salton and McGill, 1986), where the size of vectors is equal to that of the document vocabulary. We adapt the centroid-based method introducing a distributed representation of words where each word in a document is represented by a vector of real numbers of an established size. Formally, given a corpus of documents $[D_1, D_2, \dots]$ and its vocabulary V with size $N = |V|$, we define a matrix $E \in \mathbb{R}^{N,k}$, so-called *lookup table*, where the i-th row is a word embedding of size k, $k << N$,

[1] Commonly called *word2vec*.

[2] In this section we refer to a single document, but the method can be extended to a cluster of documents.

arcade	**donkey**	**kong**	**game**	**nintendo**	**coleco**	*centroid embedding*
arcades	goat	hong	gameplay	mario	intellivision	nes
pac-man	pig	macao	multiplayer	wii	atari	gamecube
console	monkey	fung	videogame	console	nes	konami
famicom	horse	taiwan	rpg	nes	msx	wii
sega	cow	wong	gamespot	gamecube	3do	famicom

Table 1: Centroid words of the *Donkey Kong (video game)* article having the tf-idf values greater than a topic threshold equal to 0.3. For each centroid word, the five closest words are shown using the skip-gram model trained on the Wikipedia (en) content. In the last column the words most similar to the centroid embedding computed using element-wise addition are shown.

of the i-th word in V. The values of the word embeddings matrix E are learned using the neural network model introduced by (Mikolov et al., 2013b). The model can be trained on the collection of documents to be summarized or on a larger corpus. This is a peculiar advantage of Representation Learning (RL) (Bengio et al., 2013) that allows to reuse an external knowledge and this is especially useful for the summarization of documents in specific domains, where large amount of data are not available. After learning the lookup table, our summarization method consists of four steps: 1) preprocess the input document; 2) build the centroid embedding; 3) compute the sentence scores; 4) select the relevant sentences.

3.1 Preprocessing

The first step follows the common pipeline for the summarization task: split the document into sentences, convert all words in lower case and remove stopwords. Stemming is not performed because we let the word embeddings to discover the linguistic regularities of words with the same root (Mikolov et al., 2013c). For instance, the most similar embeddings to the words that compose the centroid vector using the skip-gram model trained on the Wikipedia content are reported in Table 1. The closest word of *arcade* is its plural *arcades*, while they are orthogonal in the vector space according to the BOW representation.

3.2 Centroid Embedding

In order to build a centroid vector using word embeddings, we first select the meaningful words into the document. For simplicity and a fair comparison with the original method, we select those words having the $tf * idf$ weight greater than a topic threshold. Thus, we compute the centroid embedding as the sum[3] of the embeddings of the top ranked words in the document using the lookup table E.

$$C = \sum_{w \in D, tfidf(w) > t} E[idx(w)] \tag{1}$$

In the eq. (1) we denote with C the centroid embedding related to the document D and with $idx(w)$ a function that returns the index of the word w in the vocabulary. In the headers of the Table 1 the centroid words extracted from a Wikipedia article are reported. The last column shows the words most similar to centroid embedding computed using element-wise addition. It is important to underline that all five closest words to the centroid vector are semantically related to the main topic of the document despite the size of the Wikipedia vocabulary (about 1 million words).

3.3 Sentence Scoring

For each sentence in the document, we create an embedding representation by summing the vectors for each word in the sentence stored in the lookup table E.

$$S_j = \sum_{w \in S_j} E[idx(w)] \tag{2}$$

In the eq. (2) we denote with S_j the j-th sentence in the document D. Then, the sentence score is computed as the cosine similarity between the embedding of the sentence S_j and that of the centroid C of the document D.

$$sim(C, S_j) = \frac{C^T \bullet S_j}{||C|| \cdot ||S_j||} \tag{3}$$

Figure 1 shows a visualization of sentence and centroid embeddings of a Wikipedia article. We

[3] Some works use the average rather than the addition to compose word embeddings. However, the sum and the average do not change the similarity value when using the cosine distance, since the angle between vectors remains the same.

Sent ID	Sentence	Score
136	The original **arcade** version of the **game** appears in the **Nintendo** 64 **game Donkey Kong** 64.	0.9533
131	The **game** was ported to **Nintendo**'s Family Computer (*Famicom*) *console* in 1983 as one of the system's three launch titles; the same version was a launch title for the *Famicom*'s North American version, the **Nintendo** Entertainment System (*NES*).	0.9375
186	In 2004, **Nintendo** released *Mario* vs. **Donkey Kong**, a sequel to the **Game** Boy title.	0.9366
192	In 2007, **Donkey Kong** Barrel Blast was released for the **Nintendo** *Wii*.	0.9362
135	The *NES* version was re-released as an unlockable game in Animal Crossing for the *GameCube* and as an item for purchase on the *Wii*'s Virtual *Console*.	0.9308

Table 2: The most relevant sentences of the *Donkey Kong* article selected with the centroid-based summarization method using word embeddings. For each sentence are reported the related position ID in the document and the similarity score computed between sentence and centroid embeddings. The words that compose the centroid vector are marked in **bold**. The most similar words to the centroid ones are reported in *italic*.

use t-SNE method (van der Maaten and Hinton, 2008) to reduce the dimensionality of vectors from 300 to 2. For each sentence the position ID in the document is shown. The closest sentences to the centroid embedding are marked in green. The words that compose the centroid are the same showed in Table 1. In Table 2 we report the sentences near to the centroid with the related cosine similarity values. As we expected, the most relevant sentence (136) contains many words close to the centroid vector. However, the relevant aspect concerns the last sentence (135). Despite this sentence does not contain any centroid word, it has a high similarity value so it is close to the centroid embedding in the vector space. The reason is due to the presence of the words, such as *NES*, *GameCube* and *Wii*, that are the closest words to the centroid embedding (Table 1). This proves the effectiveness of the compositionality of word embeddings to encode the semantic relations between words through vector dense representations.

3.4 Sentence Selection

The sentences are sorted in descending order of their similarity scores. The top ranked sentences are iteratively selected and added to the summary until the limit[4] is reached. In order to satisfy the redundancy property, during the iteration we compute the cosine similarity between the next sentence and each one already in the summary. We discard the incoming sentence if the similarity value is greater than a threshold. This procedure is reported in Algorithm 1. However, similar sentence selection approaches are described in (Carbonell and Goldstein, 1998; Saggion and Gaizauskas, 2004).

```
Algorithm 1 Sentence selection
Input: $S$, $Scores$, $st$, $limit$
Output: $Summary$
  $S \leftarrow$ SORTDESC($S$,$Scores$)
  $k \leftarrow 1$
  for $i \leftarrow 1$ to $m$ do
      $length \leftarrow$ LENGTH($Summary$)
      if $length > limit$ then return $Summary$
      $SV \leftarrow$ SUMVECTORS($S[i]$)
      $include \leftarrow True$
      for $j \leftarrow 1$ to $k$ do
          $SV2 \leftarrow$ SUMVECTORS($Summary[j]$)
          $sim \leftarrow$ SIMILARITY($SV$,$SV2$)
          if $sim > st$ then
              $include \leftarrow False$
      if $include$ then
          $Summary[k] \leftarrow S[i]$
          $k \leftarrow k + 1$
```

4 Experiments

In this section we describe the benchmarks conducted on two text summarization tasks. The main goal is to compare the centroid-based method using two different representations (bag-of-words and word embeddings). In Section 4.1 and in Section 4.2 we report the experimental results carried out on Multi-Document and Multilingual Single Document summarization tasks, respectively.

[4] The limit can be the number of bytes/words in the summary or a compression rate.

4.1 Multi-Document Summarization

Dataset and Metrics During the document understanding conference (DUC)[5] from 2001 to 2007, several gold standard datasets have been developed to evaluate the summarization methods. In particular, we evaluate our method on multi-document summarization using the DUC-2004 Task 2 dataset composed by 50 clusters, each of which consists of 10 documents coming from Associated Press and New York Times newswires. For each cluster, four summaries written by different humans were supplied. For the evaluation, we adopt the ROUGE (Lin, 2004), a set of recall-based metrics that compare the automatic and human summaries on the basis of the n-gram overlap. In our experiment, we adopt both ROUGE-1 and ROUGE-2[6].

Baselines For the comparison, we propose several baselines. Firstly, we adapt the centroid method proposed by (Radev et al., 2004) (**C_BOW**) for a fair comparison. In the original work, the sentence scores are the linear combination of the centroid score, the positional value and the first sentence overlap. The centroid score is the sum of $tf * idf$ weights of the words occurring both in the sentence and in the centroid. In our experiment, we apply both our sentence score and selection algorithms. **LEAD** simply chooses the first 665 bytes from the most recent article in each cluster. **SumBasic** is a simple probabilistic method proposed by (Nenkova and Vanderwende, 2005) commonly used as baseline in the summarization evaluation. **Peer65** is the winning system in DUC-2004 Task 2. To compare our method with others which also use compact and dense representations, we use the method proposed by (Lee et al., 2009) that adopts the generic relevance of sentences method using **NMF**. Another method often used in summarization evaluations is **LexRank** proposed by (Erkan and Radev, 2004) which uses the TextRank algorithm (Mihalcea and Tarau, 2004) to establish a ranking between sentences. Finally, we compare our method with the one proposed by (Cao et al., 2015) that uses Recursive Neural Network (**RNN**) for learning sentence embeddings by encoding syntactic features.

System	R1	R2	tt	st	size
LEAD	32.42	6.42			
SumBasic	37.27	8.58			
Peer65	38.22	9.18			
NMF	31.60	6.31			
LexRank	37.58	8.78			
RNN	38.78	9.86			
C_BOW	37.76	8.08	0.1	0.6	
C_GNEWS	37.91	8.45	0.2	0.9	300
C_CBOW	38.68	8.93	0.3	0.93	200
C_SKIP	**38.81**	**9.97**	0.3	0.94	400

Table 3: ROUGE scores (%) on DUC-2004 dataset. **tt** and **st** are the topic and similarity thresholds respectively. **size** is the dimension of embeddings.

Implementation Our system[7] is written in Python by relying on *nltk*, *scikit-learn* and *gensim* libraries for text preprocessing, building the sentence-term matrix and import the word2vec model. We train the word embeddings on DUC-2004 corpus using the original word2vec[8] implementation. We test both continuous bag-of-words (**C_CBOW**) and skip-gram (**C_SKIP**) neural architectures proposed in (Mikolov et al., 2013a) using the same parameters[9] but varying the embedding sizes. Moreover, we compare our method using the model trained on a part of Google News dataset (**C_GNEWS**) which consists of about 100 billion words. In the preprocessing step each cluster of documents is divided in sentences and stopwords are removed. We do not perform stemming as reported in Section 3. To find the best parameters configuration, we run a grid search using this setting: embedding size in [100, 200, 300, 400, 500], topic and similarity thresholds respectively in [0, 0.5] and [0.5, 1] with a step of 0.01.

Results and Discussion The results of the experiment are shown in Table 3. We report the best scores of our method using the three different word2vec models along with their parameters. For all word embeddings models, our method outperforms the original centroid one. In detail, with the skip-gram model we obtain an increment of 1.05% and 1.71% with respect to the BOW model using ROUGE-1 and ROUGE-2 respectively. Moreover, our simple method with skip-gram performs bet-

[5] http://duc.nist.gov

[6] ROUGE-1.5.5 with options -c 95 -b 665 -m -n 2 -x

[7] https://github.com/gaetangate/text-summarizer

[8] https://code.google.com/archive/p/word2vec/

[9] -hs 1 -min-count 0 -window 10 -negative 10 -iter 10

ter than the more complex models based on RNN. This proves the effectiveness of the compositional capability of word embeddings in order to encode the information word sequences by applying a simple sum of word vectors, as already proved in (Wieting et al., 2015). Although our method with the model pre-trained on Google News does not achieve the best score, it is interesting to notice the flexibility of the word embeddings in reusing external knowledge. Regarding the comparison between BOW and embedding representations, the experiment shows different behaviors of the similarity threshold. In particular, the use of word2vec requires a higher threshold because the word embeddings are dense vectors unlike the sparse representation of BOW. This proves that the embeddings of sentences are closer in the vector space, thus the cosine similarity returns closer values.

System	R1	R2	tt	st	size
C_BOW	37.56	**8.26**	0	0.6	
C_GNEWS	36.91	7.35	0	0.9	300
C_CBOW	**37.69**	7.64	0	0.83	300
C_SKIP	37.61	8.10	0	0.91	300

Table 4: ROUGE scores without topic threshold.

Also the topic threshold shows different trends. The word embeddings require a higher threshold value to make our method effective. In order to analyze this aspect we run another experiment setting the topic threshold to 0. The results are reported in Table 4. Results show that the BOW representation is more stable and obtains the best ROUGE-2 score, while the performance obtained by word2vec decreases considerably. This means that word embeddings are more sensitive to noise and they require an accurate choice of the meaningful words to compose the centroid vector.

Summaries Overlap Although the different methods achieve similar ROUGE scores, they not necessarily generate similar summaries. An example is reported in Table 6. In this section we conduct a further analysis by comparing the summaries generated by the best four configurations of the centroid method reported in Table 3. We adopt the same criterion presented in (Hong et al., 2014), where the different summaries are compared in terms of sentences and words overlap using the Jaccard coefficient. Due to space constraint, we report in Table 5 only the sentence overlap. The results prove that different word representations lead to different summaries. In particular, the summaries using BOW differ considerably from those generated using word2vec, but this is true even for different embedding models. On the other hand, only the models trained on the DUC-2004 corpus (CBOW and SKIP) tend to generate more similar summaries. This analysis suggests that a combination of various models trained on different corpora could result in good performance.

	GNEWS	CBOW	SKIP	BOW
GNEWS	1	0.109	0.171	0.075
CBOW		1	0.460	0.072
SKIP			1	0.105
BOW				1

Table 5: Sentence overlap.

4.2 Multilingual Document Summarization

Task Description We carried out an experiment on Multilingual Single-document Summarization (MSS). Our main goal is to prove empirically the effectiveness of the use of word embeddings in the document summarization task across different languages. For this purpose, we evaluate our method on the MSS task proposed in MultiLing 2015 (Giannakopoulos et al., 2015), a special session at SIGDIAL 2015. Starting from 2011 the aim of MultiLing community is to promote the cutting-edge research in automatic summarization by providing datasets and by introducing several pilot tasks to encourage further developments in single and multi-document summarization and in summarizing human dialogs in on-line forums and customer call centers. The goal of the MSS 2015 task is to generate a single document summary from a selection of some of the best written Wikipedia articles with at least one out of 38 languages defined by organizers of the task. The dataset[10] is divided into a training and a test sets, both consisting of 30 documents for each of 38 languages. For both datasets, the body of the articles and the related abstracts with the character length limits are provided. Since the Wikipedia abstracts are summaries written by humans, they are useful to perform automatic evaluations. We evaluate our method using five different languages: English, Italian, German, Spanish and French.

[10] http://multiling.iit.demokritos.gr/pages/view/1532/task-mss-single-document-summarization-data-and-information

BOW - Bag of Words baseline
The controversy centers on the payment of nearly dlrs 400,000 in scholarships to relatives of IOC members by the Salt Lake bid committee which won the right to stage the 2002 games. Pound said the panel would investigate allegations that "there may or may not have been payments for the benefit of members of the IOC or their families connected with the Salt Lake City bid." Samaranch said he was surprised at the allegations of corruption in the International Olympic Committee made by senior Swiss member Marc Hodler.
CBOW - Continuous Bag of Words trained on DUC-2004 dataset
Marc Hodler, a senior member of the International Olympic Committee executive board, alleged malpractices in the voting for the 1996 Atlanta Games, 2000 Sydney Olympics and 2002 Salt Lake Games. The IOC, meanwhile, said it was prepared to investigate allegations made by Hodler of bribery in the selection of Olympic host cities. The issue of vote-buying came to the fore in Lausanne because of the recent disclosure of scholarship payments made to six relatives of IOC members by Salt Lake City officials during their successful bid to play host to the 2002 Winter Games.
SKIP - Skip-gram trained on DUC-2004 dataset
Marc Hodler, a senior member of the International Olympic Committee executive board, alleged malpractices in the voting for the 1996 Atlanta Games, 2000 Sydney Olympics and 2002 Salt Lake Games. The IOC, meanwhile, said it was prepared to investigate allegations made by Hodler of bribery in the selection of Olympic host cities. Saying "if we have to clean, we will clean," Juan Antonio Samaranch responded on Sunday to allegations of corruption in the Olympic bidding process by declaring that IOC members who were found to have accepted bribes from candidate cities could be expelled.
GNEWS - Skip-gram trained on Google News dataset
The International Olympic Committee has ordered a top-level investigation into the payment of nearly dlrs 400,000 in scholarships to relatives of IOC members by the Salt Lake group which won the bid for the 2002 Winter Games. The mayor of the Japanese city of Nagano, site of the 1998 Winter Olympics, denied allegations that city officials bribed members of the International Olympic Committee to win the right to host the games. Swiss IOC executive board member Marc Hodler said Sunday he might be thrown out of the International Olympic Committee for making allegations of corruption within the Olympic movement.

Table 6: Summaries of the cluster **d30038** in DUC-2004 dataset using the centroid-based summarization method with different sentence representations.

Model Configuration In order to learn word embeddings for the different languages, we exploit five Wikipedia dumps[11], one for each chosen language. We extract the plain text from the Wiki markup language using Wikiextractor[12], a Wikimedia parser written in Python. Each article is converted from UTF-8 to ASCII encoding using the Unidecode Python package. Since in the previous evaluation we observe a similar behavior between the continuous bag of words and skip-gram models, in this evaluation we adopt only the skip-gram one using the same training parameters[13] for all five languages. The Table 7 reports the Wikipedia statistics for the five languages regarding the number of words and the size of the vocabularies.

Language	# Words	Vocabulary
English	1,890,356,976	973,839
Italian	371,218,773	378,286
German	657,234,125	1,042,683
Spanish	464,465,399	419,683
French	551,057,299	458,748

Table 7: Wikipedia statistics.

Experiment Protocol In order to reproduce the same challenging scenario of the MultiLing 2015 MSS task, we performed the tuning of parameters using only the training set. To find the best topic and similarity threshold parameters we run a grid search as explained in Section 4.1. The grid search is performed for each language separately using both BOW and skip-gram representations. The parameter configurations are in line with those of the previous experiment on DUC-2004. In detail, the topic thresholds are in the range [0.1, 0.2] using the BOW model and in the range [0.3, 0.5] using word embeddings. While, the similarity thresholds are slightly higher w.r.t. the multi-document experiment: about 0.7 and 0.95 for BOW and skip-gram, respectively. This is due to the fact that too similar sentences are rare, especially with well-written documents as Wikipedia articles. The best parameters configuration for each language is used to generate summaries for the documents in the test set. Also for this task, each document is preprocessed with the sentences segmentation and stopwords removal, without stemming. We adopt the same automatic evaluation metrics used by the participating systems in MSS 2015 task: ROUGE-1, -2, -SU4[14]. ROUGE-SU4 computes the score between the generated and human summaries considering the overlap of the skip-bigrams of 4 as well as the unigrams. Finally, the generated summary for each document must comply with a specific length constraint (rather than using a unique length limit for the whole collection). This differs

[11] https://dumps.wikimedia.org/_*lang*_wiki/20161220/ with _*lang*_ in [en, it, de, es, fr]

[12] https://github.com/attardi/wikiextractor/wiki

[13] -hs 1 -min-count 10 -window 8 -negative 5 -iter 5

[14] ROUGE-1.5.7 with options -n 2 -2 4 -u -x -m

	English		Italian		German		Spanish		French	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
LEAD	44.33	11.68	30.46	4.38	29.13	3.21	43.02	9.17	42.73	8.07
WORST	37.17	9.93	39.68	10.01	33.02	4.88	45.20	13.04	46.68	12.96
BEST	50.38	15.10	43.87	12.50	40.58	8.80	53.23	17.86	51.39	15.38
C_BOW	49.06	13.43	33.44	4.82	35.28	4.93	48.38	12.88	46.13	10.45
C_W2V	**50.43**‡	13.34†	**35.12**	**6.81**	**35.38**†	**5.39**†	**49.25**†	**12.99**	**47.82**†	**12.15**
ORACLE	61.91	22.42	53.31	17.51	54.34	13.32	62.55	22.36	58.68	17.18

Table 8: ROUGE-1, -2 scores (%) on MultiLing MSS 2015 dataset for five different languages.

from the previous evaluation on DUC-2004.

Results and Discussion The results for each languages are shown in Table 8. We report the ROUGE-1, -2 scores for each chosen language. **LEAD** and **C_BOW** represent the same baselines used in the multi-document experiment. The former uses the initial text of each article truncated to the length of the Wikipedia abstract. The latter is the centroid-based method with the BOW representation. Our method that uses word embeddings learned with skip-gram model is labeled with **C_W2V**. For each metric and language we also report the **WORST** and the **BEST** scores[15] obtained by the 23 participating systems at MSS 2015 task. Finally, **ORACLE** scores can be considered as an upper bound approximation for the extractive summarization methods. It uses a covering algorithm (Davis et al., 2012) that selects sentences from the original text covering the words in the summary without disregarding the length limit. We highlight in bold the scores of our method when it outperforms the baseline **C_BOW**. On the other hand, the superscripts † and ‡ imply a better performance of our method with respect to the **WORST** and the **BEST** scores respectively.

Both centroid-based methods overcome the simple baseline over all languages. Our method always achieves better scores against the BOW model except for ROUGE-2 metric for English. This confirms the effectiveness of using word embeddings as alternative sentence representations able to capture the semantic similarities between the centroid words and the sentences, when summarizing single documents too. Moreover, our method outperforms substantially the lowest scores performing systems participating in MSS 2015 task for English and German languages. For English our method obtains a ROUGE-1 score even better than the one of the best system in MSS 2015. Instead, our method fails in summarizing Italian documents and it achieves the worst ROUGE-2 for Spanish and French languages. The reason may lie in the size of the Wikipedia dumps used to learn the word embeddings for different languages. As showing in Table 7, the sizes of the various corpora as well as the ratios between the number of words and dimension of the vocabularies, differ consistently. The English version of Wikipedia consists of nearly 2 billion words against about 300 million words of Italian one. Thus, according to the distributional hypothesis reported in (Harris, 1954), we expect better performance for our method in summarizing English or German articles with respect to the other languages where the word embeddings are learned using a smaller corpus. Our results and in particular the ROUGE-SU4 scores reported in Figure 2 support this hypothesis.

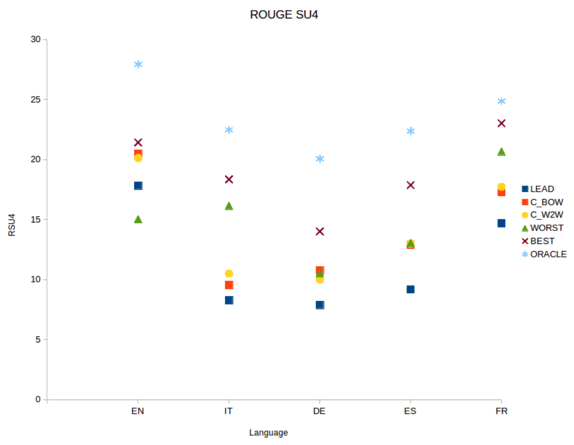

Figure 2: ROUGE-SU4 scores (%) comparison on MultiLing MSS 2015 dataset.

[15] http://multiling.iit.demokritos.gr/file/view/1629/mss-multilingual-single-document-summarization-submission-and-automatic-score

5 Conclusion

In this paper, we propose a centroid-based method for extractive summarization which exploits the compositional capability of word embeddings. One of the advantages of our method lies on its simplicity. Indeed, it can be used as a baseline in experimenting new articulate semantic representations in summarization tasks. Moreover, following the idea of representation learning, it is feasible to infuse knowledge by training the word embeddings from external sources. Finally, the proposed method is fully unsupervised, thus it can be adopted in other summarization tasks, such as query-based document summarization. As future work, we plan to evaluate the centroid-based summarization method using a topic model, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) or Non-negative Matrix Factorization (NMF) (Berry et al., 2007), in order to extract the meaningful words to compute the centroid embedding as well as to carry out a comprehensive comparison of different sentence representations using more complex neural language models (Le and Mikolov, 2014; Zhang and LeCun, 2015; Józefowicz et al., 2016). Finally, the combination of distributional and relational semantics (Fried and Duh, 2014; Verga and McCallum, 2016; Rossiello, 2016) applied to extractive text summarization is a promising further direction that we want to investigate.

References

Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, Aug.

Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, September.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2153–2159. AAAI Press.

Jaime Carbonell and Jade Goldstein. 1998. 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, New York, NY, USA. ACM.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. pages 484–494, August.

S. T. Davis, J. M. Conroy, and J. D. Schlesinger. 2012. Occams – an optimal combinatorial covering algorithm for multi-document summarization. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 454–463, Dec.

Gunes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, December.

Daniel Fried and Kevin Duh. 2014. Incorporating both distributional and relational semantics in word representations. *CoRR*, abs/1412.4369.

George Giannakopoulos, Jeff Kubina, John M. Conroy, Josef Steinberger, Benoît Favre, Mijail A. Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. Multiling 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the SIGDIAL 2015 Conference.*, pages 270–274.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. 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.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196. JMLR Workshop and Conference Proceedings.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521:436–444.

Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. 2009. Automatic generic document summarization based on non-negative matrix factorization. *Inf. Process. Manage.*, 45(1):20–34, January.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. pages 280–290, August.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101.*

Makbule G. Ozsoy, Ferda N. Alpaslan, and Ilyas Cicekli. 2011. Text summarization using latent semantic analysis. *Journal of Information Science*, 37(4):405–417.

Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938, November.

Gaetano Rossiello. 2016. Neural abstractive text summarization. In *Proceedings of the Doctoral Consortium of AI*IA 2016 co-located with the 15th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2016), Genova, Italy, November 29, 2016.*, pages 70–75.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September. Association for Computational Linguistics.

Horacio Saggion and Robert Gaizauskas. 2004. Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the HLT/NAACL Document Understanding Workshop (DUC 2004)*, Boston, May.

Horacio Saggion and Thierry Poibeau. 2013. Automatic text summarization: Past, present and future. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 3–21. Springer.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Patrick Verga and Andrew McCallum. 2016. Row-less universal schema. *CoRR*, abs/1604.06361.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *CoRR*, abs/1511.08198.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *CoRR*, abs/1502.01710.