# Automatic Text Summarization by Non-topic Relevance Estimation

Ignacio Arroyo-Fernández[1,2], Juan-Manuel Torres-Moreno[2], Gerardo Sierra[1]
and Luis Adrián Cabrera-Diego[2]
[1]*Institute of Engineering, UNAM, Mexico City, Mexico*
[2]*Laboratoire Informatique d'Avignon, UAPV, Avignon, France*

Keywords: Automatic Text Summarization, Machine Learning, Generalization Ability, Regression Estimation.

Abstract: We investigate a novel framework for Automatic Text Summarization. In this framework underlying language-use features are learned from a minimal sample corpus. We argue the low complexity of this kind of features allows relying in generalization ability of a learning machine, rather than in diverse human-abstracted summaries. In this way, our method reliably estimates a relevance measure for predicting summary candidature scores, regardless topics in unseen documents. Our output summaries are comparable to the state-of-the-art. Thus we show that in order to extract meaning summaries, it is not crucial what is being said; but rather how it is being said.

## 1 INTRODUCTION

Automatic Text Summarization (ATS) is a Natural Language Processing (NLP) technique, which aims to filter relevant content from redundant one. However, currently it is difficult to determine what a *"good summary"* actually is. Due to this difficulty ATS can become into a very subjective task, even when human-generated summaries are attempted to be evaluated.

Most state-of-the-art ATS methods rely on the maximum intersection of lexical features in source document(s)/sentence(s). Sometimes human-generated *(abstractive/extractive) summaries* are used as training data (Kupiec et al., 1995; Nenkova and Passonneau, 2004; Torres-Moreno, 2014). This approach leads the summarizer to be focused in topic features the documents specifically contain (words, terms, entities, etc.) Likewise, well known ATS evaluation methods take advantage of human-generated summaries (if they are available), which are used as comparison points (Lin, 2004). Currently this outline holds a reasonable support in state-of-the-art *extractive summarization* (Louis and Nenkova, 2009), which is of our interest in this paper.

Despite of their reliability, these approaches suffer from high sensibility to the size of the source documents/sentences. This scenario can get worse in situations where topics are too specific, where the vocabulary changes or where human summaries on such a specific topic are not available for training. For instance, see the two following sentences:

> *"Given that we are hosting the event, we strongly recommend you that between now and early Monday, to make cleaning your space in a meticulous way, that is, not having dirty cups or papers strewn everywhere."* (1)

> *"The arrangement of genes from telomere to centromere is G11-C4A-ZA-21A-YA-XA-C4B-ZB-21B-YB-XBS-XB."* (2)

These sentences are very different in terms of simple language-use observations, regardless the topic they deal with. On one hand, it can be observed that (1) is relatively too large and it has many different words (this looks complex at first glace). Nonetheless, the vocabulary is very usual and majority of words are certainly redundant for the conveyed message/topic (which seems to be underlying and noisy, rather than to be clear: *"Keep clean your place."*). On the other hand, (2) is relatively short and denotes specific key information about the directional manner a pair of `specific nouns` are related by means of another much more `specific noun`. Notice that the last explaining sentence is too general and that it has free arguments: two `specific nouns` (*"telomere"* and *"centromere"*) and `what relates` them (the *"arrangement of genes G11-C4A-ZA-21A-YA-XA-C4B-ZB-21B-YB-XBS-XB"*).

Arroyo-Fernández, I., Torres-Moreno, J-M., Sierra, G. and Cabrera-Diego, L.
Automatic Text Summarization by Non-topic Relevance Estimation.
DOI: 10.5220/0006053400890100
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 89-100
ISBN: 978-989-758-203-5

The above motivates us to propose that a good summary not necessarily depends on the maximum lexicon intersection, but also (and probably more importantly) on how words/terms are used. In (Harris, 1968), it is indeed stated that word combinatorial constraints (including redundancy) lead to information perception in context. In an attempt to be more akin to these concepts, we present an initial study for leveraging *language-use features* as dominant and underlying cues for sentence relevance estimation.

Given that we are analyzing basic principles of the ATS task, we present *non-topic-oriented* extractive ATS as a general framework (which does not require to identify topics). As a simple analysis approach, we used a learning machine to filter underlying language-use features from *Doc2Vec* (D2V) *Paragraph Vector* representations (sentence embeddings in our case) (Le and Mikolov, 2014). Notice that D2V provides general and unstructured sentence representations from unstructured text, which is a potential source of noise[1]. These representations are inferred from a tiny training set.

We see this training set as a minimal sample (a human-generated extractive summary of only 30 sentences), which is motivated by the supposition of that language-use features are much less complex[2] than topic ones. Thus a SVR (Support Vector Regression) machine learned the relationship between each training sentence and its associated summary candidature score.

Acquired knowledge from our minimal sample provides impressive generalization performance over a much greater test corpus, where multiple documents include a wide variety of topics (which are not in the training set). The Rouge and The Fresa ATS evaluation measures were used to validate the mentioned performance (Lin, 2004; Saggion et al., 2010).

Along with summary evaluations, we show simple statistics for comparing concentration of state-of-the-art features in machine summaries against their concentration in human references. Likewise, the same comparison was performed for some other features that we think that are very important, but less considered in the state of the art (Min et al., 2012; Li et al., 2013; Hong and Nenkova, 2014). In this way, we show that for keeping contents it is not necessary to pay primary attention to lexicon.

The structure of the paper is as follows: Section 2 shows the related work, Section 3 describes our datasets, Sections 4 and 5 expose our learning algorithms, Section 7 describes our experimental setup, Section 8 explains our results, Section 9 presents a detailed discussion and finally Section 10 gives our conclusion.

## 2 RELATED WORK

Ever since the pioneering work of (Luhn, 1958), ATS techniques have experienced evolution in a bast variety of forms. The main and well known categorization of techniques divides them into single-document and multi-document ATS ones.

In both categories the output summary must be substantially smaller than the input source documents, which supposes a compression rate (Torres-Moreno, 2014). Herein we are particularly interested in ATS frameworks taking advantage of machine learning techniques, where human references are used for training.

The idea of using machine learning techniques in ATS is relatively new. Most contributions rely on the assumption of that word counting is fundamental. The seminal contribution by (Kupiec et al., 1995) uses the Bayes theorem for classifying two classes of sentences: those that are relevant and those that are not. The corpus the author presented is constituted of training pairs (`source_document,human_reference`), where such pairs are treated as a kind of association between an input sample (the source document) and its example label (the human reference). Intuitive feature engineering is used to compute a statistical relationship between all elements of the set of training pairs (`source_document,human_reference`), which includes POS tags, POS *n*-grams, *topic signatures*[3] (e.g. unigrams, bigrams, named entities, etc.), sentence size, sentence position, etc. Overall, the above idea remains nowadays. Nonetheless, one observable difference is the utilization of varied and well known statistical data analysis methods, e.g. matrix decomposition and graph algorithms (Landauer et al., 1998; Lee et al., 2009; Divya and Reghuraj, 2014).

Traditionally statistical analysis methods operate over some combination of the abovementioned features, which are represented in term-document or graph matrices. The main purpose of these techniques is to filter those source words/terms/sentences whose

[1] In general, we consider that any vector representation method has many potential sources of noise and of computational cost, whenever redundant data is not properly filtered (e.g. when a complete parse tree is used for ATS.)

[2] Complexity in the frame of sampling can be understood as the minimum amount of different cues can satisfactorily explain a given set of patterns.

[3] Topic signatures can be considered as particular cases of topic features in general.

linear variance is maximum with respect to the rest of them. In this way, a ranking is performed and relevant sentences for summary are picked up according to a user-defined threshold (or compression rate).

Both in single-document and multi-document summarization methods, redundancy has been an important issue. In multi-document scenarios source documents not only deal with different subtopics, but also (and majorly) with shared subtopics. In an effort of discriminating redundancy caused by shared subtopics, text clustering techniques have been employed (Wan and Yang, 2008; Sarkar, 2009). Clusters of subtopics are built inside documents, then a by-cluster ranking is performed in order to obtain a global ranking of relevant summaries.

There are more specific contributions which employ machine learning methods like the used one in this work. In these cases supervised learning was used for classification/regression estimation of semi-supervised scores (Min et al., 2012; Li et al., 2013). In these contributions, varied settings of feature engineering take place, e.g. *n*-gram sharing, word sharing, named entities and knowledge base similarities (e.g. the Lin similarity used in WordNet).

The main differences among these methods is the procedure for generating summary candidature scores. In the case of (Li et al., 2007), ranking scores were actually the lexical semantic similarities between sentences in source documents and sentences in human references. A similar approach was followed in (Galanis et al., 2012), but they performed regression estimation over the average between Rouge-2 and Rouge-4 measures. In the case of (Berg-Kirkpatrick et al., 2011; McDonald, 2007), a linear kernel SVM was used to separate bad summary sentences from good ones. The authors constructed vectors from weights of document bigrams. These weights were derived from how bigrams were shared between source documents and human references.

Recent contributions used structural features in their methodologies. The method proposed by (Woodsend and Lapata, 2012) has the ability of being extractive and abstractive. The authors computed the importance of both words and bigrams by means of structural features acquired from the complete parse tree of source sentences. Thus, similarly to (McDonald, 2007), vector representations from *n*-gram counts and structural features are constructed. These vectors are separated by a linear SVM in order to select composing sentences of the output summary.

More recently, (Cao et al., 2015) appeals to reduction in feature engineering by using minimally hand-crafted “word features”. A recursive neural network was trained over the complete parse trees of source sentences and human summaries. A hierarchical regression problem was formulated over sentence salience in the tree paths. Once the network is trained, the importance of source sentences was predicted over the parsed test sentences.

In our proposal we are using characteristics of the two main ATS categories, i.e. our training framework is single-document and our test framework is multi-document. As our investigated framework is of a general character, we considered recent advances in vector representation of text (Mikolov et al., 2013a; Le and Mikolov, 2014), as well as basic properties of kernel machines (Schölkopf et al., 1998). The main advantage of these machines is the control we can have over the learning procedure and their suitability in low sampling scenarios. With these elements in mind, we use the mentioned advantage to propose a novel ATS framework. This framework relies in generalization predictability of the learning machine and classical sampling theories (Nyquist, 1924; Shannon, 1949; Cortes, 1995; Cortes and Vapnik, 1995; Vapnik, 1998; Arroyo-Fernández, 2015), rather than in bast amounts of human-generated summaries.

# 3 DATASETS

In this work we decided to process three French language datasets.

As **context corpus** for training the sentence embeddings we used the French Wikipedia, which contents range up to 2012.

As **training data** we used a dataset called Puces[4]. This is a document that concentrates relevance votes of 172 individuals for each of its 30 sentences. The amount of votes was averaged and normalized to 1 for each sentence. These average votes were used as sentence summary candidature scores. This document deals with 2 topics. One of them is content in the first 15 sentences and the other one in the second 15 ones. Each topic refers to two different senses of the polysemic French word *“puces”* (fleas).

As **test data** we employed the RPM2[5] corpus. This corpus contains two groups of news documents. Each group contains 20 topics, in such a way each topic is comprised of 10 independently generated documents. Each document in turn is associated to 4 independently abstracted human references.

Although Puces is a tiny summary sample (single-document), it was assessed by 172 different Computer Science students (both from BSc and from MSc programs). In contrast, RPM2 has multiple documents

[4] http://dev.termwatch.es/∼fresa/CORPUS/PUCES/

[5] http://rpm2.org

and uniquely 4 different human references by document. We used these references in our ATS evaluation framework.

## 4 THE SENTENCE VECTOR REPRESENTATION METHOD

Neural word embeddings provided by Word2Vec (W2V) are used as basis for building state-of-the-art sentence vector representations, i.e. D2V sentence embeddings (Mikolov et al., 2013a; Le and Mikolov, 2014). This sentence embedding technique has shown high performance in a number of NLP tasks (mainly in semantic assessments (Li et al., 2016)).

Currently, there are not reliable explanations about linguistic concepts encoded in such sentence embeddings. Nonetheless, the stochastic nature that motivated the authors for designing such an algorithm provides sufficient insight on the usefulness of its context representation properties.

The core principles governing W2V word embeddings as constituent elements of sentences are described as follows. Let $V$ be the vocabulary of a context corpus $D$, the main idea behind W2V is, firstly, modeling the presence of a given (center) word $w_i \in D$ in some context $c \in C$ with probability

$$p_c(w_i|w_{i-k},...w_{i+k}) = \frac{e^{v_{w_i}^\top v_c + b_C}}{\sum_{w\in V} v_w^\top v_c + b_V} \tag{3}$$

such that: $k = 1,2,...,|c|/2$. Herein $v_{w_i}, v_w \in \mathbb{R}^d$ are word embeddings of $w_i$ and each other different words $w \in V$, respectively. The vector $v_c$ is the resulting combination[6] of surrounding $2k$ word embeddings of the $w_{i\pm k}$ words in the context $c$. Prior to the training, all $v_w$ embeddings and $b_{(\cdot)}$ bias parameters are randomly initialized.

Secondly, the probability (3) is lead to be consistent by means of gradient descent maximization of the average log for each $w_i \in D$, wherever it appears in the corpus:

$$\max_{v_{w_i}, v_w, b_{(\cdot)}} \frac{1}{|D|} \sum_{w_i \in D} \log p_t(w_i|w_{i-k},...w_{i+k}) \tag{4}$$

Once (4) is maximum for some parameters $v_{w_i}, v_w, b_{(\cdot)}$, it can be said the model (3) is trained. It remains to show the additional mechanism that allows us to have each of sentence embeddings that we actually used in this work. Suppose in a given context $c$ we have an additional (virtual) word called $s_c$, which now denotes a sentence. This virtual word, in contrast to real ones, neither is in other context than the given one, nor is in the original $V$. The same operation is performed for all $s_c \in S_D$.

The problem of inferring a sentence embedding $v_{s_c}$ amounts to include $s_c$ in (4) as the center word and then maximize the corresponding log probability. Thus, as in the case of fitted word embeddings $v_w$, we can pick up from the set $S_D$ of sentences in $D$ the fitted sentence embedding $v_{s_c}$ from the trained model as required. The uniqueness of virtual $v_{s_c}$ through $D$ is motivated by the fact that all sentences $s_c \subset D$ are highly likely unique.

Frequently a word sequence $\{s_x \notin S_D : x \notin C\}$ takes place as an unseen sentence. Inferring such an unseen sentence limits to build the analogous embedding $v_x$. Then the prediction (inference) of the corresponding virtual embedding $v_{s_x}$ is easy. This is because the uniqueness assumption and because highly likely any $w \in s_x$ is already learned as an embedding $v_w \in v_x$ (the integrity of $v_x$ depends on $|V|$).

## 5 THE SUMMARIZER LEARNING MACHINE

Usualy semantic (or lexical) similarity between source documents and human references is used as main cue for learning state-of-the-art summarizers. Unlike to these usual approaches, in this work a learning machine is taught to keep relevant contents of a document. This happens independently of the particular topics or semantics they convey. The machine learns the relationship between sentences and their summary candidature scores from Puces.

The learned relationship takes into account more general features than those which are used to measure sentence semantic similarity. We consider not only features that make sentences good summary candidates are useful, but also those features that make sentences bad candidates do so. Thus, a regression problem was proposed. By solving such a problem, an approximated relationship between the summary candidature scores and the language-use features encoded in sentence embeddings is learned.

On the basis of that a kernel machine fits a nonlinear transformation whose input space is any vector representation (even when it does not define a vector space), we decided to use the SVR algorithm (Schölkopf et al., 1998).

Let $X$ be some input space and let $\mathcal{X} = x_1,...,x_\ell \subset X \subset \mathbb{R}^d$ be a training set of sentence embeddings, then the well known representer theorem is defined

[6] Such a combination can be simple averaging or concatenation of constituent word embeddings in $c$.

(Shawe-Taylor and Cristianini, 2004):

$$f_{\alpha^*}(x) = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_{i'}^*) k_\gamma(x, x_i) + b \qquad (5)$$

In equation (5) the estimator $f_{\alpha^*}(\cdot) \in \mathfrak{F}_\alpha$ is a linear combination, which is aimed to approximate the non-linear empirical function given by the human-judged summary candidature scores $\mathcal{Y} = y_1, ..., y_\ell \in [0,1]$. A 0 valued candidature is the lowest (for bad candidates) and an 1 valued candidature is the highest one (for the best candidates). $b \in \mathbb{R}$ is a bias parameter. The function $k_\gamma(\cdot,\cdot)$ is a positive definite kernel, which is parametrized by $\gamma$ (e.g. the bandwidth parameter for Gaussian kernels). This kernel function acts as an inner product between any $x \in X$ and each $x_i \in \mathcal{X}$. The approximation is possible by finding optimal $\alpha = \alpha_1, ..., \alpha_\ell$ coefficients (i.e. $\alpha^*$) over which the set of possible solutions $\mathfrak{F}_\alpha$ is defined. In this way, it is attempted the learning machine to approximate the correspondence $f_{\alpha^*}(\mathcal{X}) \rightarrow \mathcal{Y}$. For this purpose the structural risk functional must be minimized (Vapnik, 1998):

$$\min_{f_\alpha} \mathcal{R}_{estr}(\mathfrak{F}_\alpha) = R_{emp}(\mathfrak{F}_\alpha, \mathcal{Y}) + \lambda \phi(\mathfrak{F}_\alpha) \qquad (6)$$

where $f_\alpha \in \mathfrak{F}_\alpha$. When there exists a Hilbert space $\mathcal{H}$ such that $\mathfrak{F}_\alpha \subset \mathcal{H}$, $R_{emp}(\cdot,\cdot)$ is the error $\ell_2$ norm $\|f_\alpha(x_i) - y_i\|_2$ and $\phi(\cdot)$ is the estimator's $L_2$ norm $\|f_\alpha(x)\|_2$. $\lambda \in \mathbb{R}_+$ is the so-called regularization parameter, which controls the precision of $f_{\alpha^*}(\mathcal{X}) \rightarrow \mathcal{Y}$ and the cardinality of $\mathfrak{F}_\alpha$.

# 6 ATS EVALUATION

In this work we used two evaluation methods: Rouge and Fresa. The former uses human references and the second does not. Thus we report ATS evaluation with and without human references (Louis and Nenkova, 2009). There are also many aspects that can give insight about the quality of a summary, e.g. informativeness, coherence, precision, recall, etc. Nonetheless, this is a general study on basics of ATS, so at the moment we only consider statisical aspects related to intrinsic recall and informativeness (Cabrera-Diego et al., 2016).

## 6.1 The Rouge Measure

Rouge is the most popular among human-referenced methods in the literature (Lin, 2004). Rouge focuses on recall between the $n$-grams of machine summaries and the $n$-grams of human references. There are four commonly used Rouge measures.

The first variant (Rouge-1) measures the recall between the unigrams of the source document and those of the associated machine summary, the second and third variants (Rouge-2 and Rouge-3) perform the same measurement, but for bigrams and three-grams, respectively. The fourth variant (Rouge-SU4) also performs the same measurements than Rouge{-1-2-3}, but it skips at most four unigrams to build a bigram.

The general formula of the Rouge measure is given as follows:

$$Rouge_n = \frac{\sum_{s \in R} \sum_{g_n \in s} C_m(g_n)}{\sum_{s \in R} \sum_{g_n \in s} C(g_n)}$$

where $R$ is the set of reference summaries, $s$ is a candidate summary, $g_n$ is some $n$-gram, $C(\cdot)$ is an accumulator function, i.e. it counts how many times its argument appears in $s$. $C_m(\cdot)$ is a conditional accumulator function, i.e. it counts only when its argument matches in both, the candidate summary $s$ and the human references $R$. Thus, notice the fundamental roll the lexical intersection assumption plays here.

## 6.2 The Fresa Measure

Unlike to Rouge, Fresa evaluates summaries without using human references. To this end, the distributional divergence between sources and human references is computed, i.e. the Kullback-Leibler (KL) divergence (Saggion et al., 2010). Therefore if Fresa is defined as:

$$F = \frac{1}{n} \sum_{n \in N} D_{KL}^n(P\|Q)$$

then $n = 1, 2, 3...$ defines the KL divergences for corresponding $n$-grams (analogously to Rouge{-1-2-3}). For $n = |N|$, we have actually the same 4-skip-gram idea than for Rouge. Thus the general form of the $n$-gram KL divergence is given by:

$$D_{KL}^n(P\|Q) = \sum_{g_n \in P} \left| \log_2\left(\frac{C_P(g_n)}{|P|} + 1\right) - \log_2\left(\frac{C_Q(g_n)}{|Q|} + 1\right) \right|$$

where $P$ is the distribution of units in some machine summary and $Q$ is the distribution of units in the corresponding source document. Likewise, $C_{\{P,Q\}}(\cdot)$ is the accumulator function of the $n$-gram $g_n$ inside the subscript distribution, e.g., $P$.

## 6.3 Language-use Feature Spectrum

For each language-use feature $\mu_i$, we define its concentration in a machine summary $s$, i.e. $\varphi_{\mu_i}^s$: the number of times that $\mu_i$ appears in $s$. For human references, we have the concentrations of such a language-use feature $\mu_i$ in $h$ human references $R_1^{\mu_i},...,R_h^{\mu_i}$. Thus we defined the median[7] concentration $\varphi_{\mu_i}^R = median(R_1^{\mu_i},...,R_h^{\mu_i})$, which we used as a general human reference.

In order to compare machine summaries against our general human reference, from information theory (Shannon, 1949), we define the smoothed concentration ratio of $\varphi_{\mu_i}^s$ with respect to $\varphi_{\mu_i}^R$:

$$\mathcal{C}_{\mu_i} = 20\ \log_{10}\left(\frac{\varphi_{\mu_i}^s + 1}{\varphi_{\mu_i}^R + 1}\right) \tag{7}$$

Equation (7) is a general comparison method[8] that provides a normalized view of the difference between some reference value and any unreferenced measure. This logarithmic comparison allows that small differences are equally significant than large ones and at the same time, their magnitude orders are respected. Thus if we have a set of language-use features $\mu_1,...,\mu_n$, then its log language-use feature spectrum with respect to the general human reference (hereinafter *feature spectrum*) can be seen as the set $\mathcal{C}_{\mu_1},...,\mathcal{C}_{\mu_n} \subset \mathbb{R}$.

Notice in (7) that if $\varphi_{\mu_i}^R = \varphi_{\mu_i}^s$ then $\mathcal{C}_{\mu_i} = 0$. There are two possible reasons for this result. The first reason is that the machine summary $s$ holds not difference with respect to the human reference $R$ for the feature $\mu_i$. The second reason, is that $\mu_i$ is not informative. The case of $\mathcal{C}_{\mu_i} < 0$ means that $s$ has some lack of $\mu_i$ with respect to $R$. The case of $\mathcal{C}_{\mu_i} > 0$ means that $s$ has some excess of $\mu_i$ with respect to $R$.

# 7 EXPERIMENTAL SETUP

In this work we compare our resulting summaries against those obtained by means of the state-of-the-art Artex ATS tool (Torres-Moreno, 2012). Furthermore, we proposed some baselines for adding certainty to our analysis. From each RPM document we have three baselines: (1) *BL_rand.* The 5% of the sentences of a document were randomly picked up as summary. (2) *BL_long.* The 5% of the largest sentences of the document was picked up as summary. (3) *BL_1st.* From the first sentences of the document, the 5% was picked up as summary.

## 7.1 Sentence Vector Representations

The French Wikipedia was disposed in an unique file. Each row of this file contains a document/sentence. All non-Latin and numeric characters were removed, as well as all words were lower-cased. This preprocessing was also performed on both Puces and RPM2 datasets. The 10 documents of each of 20 topics in RPM2 were concatenated and the sentences whose length $L < 26$ were discarded.

A D2V[9] model was trained over the union between The French Wikipedia and The Puces corpus. Sentence embeddings for Puces and RPM2 were inferred as training and test vectors, respectively. All sentence embeddings (both training and test ones) were inferred from the trained D2V model. Neither sources nor human references from the RPM2 were included in the D2V training set.

Sentence embeddings of different dimensions were inferred from D2V, i.e. $d = 300,\ 200, 100, 50, 25, 20, 10, 8$ in eq. (3). The size of the context window $|c|$ was set both, to 8 and to 16 words. The following features were combined for building our experimental sentence embeddings:

- *Word-based.* The usual word based D2V sentence embeddings (Li et al., 2013).
- *Word-LP-based.* The sentence length $L$ (in amount of words) and the sentence position $P$ in the document were concatenated as two new dimensions to each of the above word-based sentence embeddings.
- *PoS-based.* These are D2V embeddings trained/inferred from PoS tags of each sentence/document in our corpus.
- *Word-PoS-based.* This is the concatenation between word-based and PoS-based sentence embeddings.

## 7.2 The SVR

The SVR was trained in a 7-fold randomized cross-validation schema. Both for the SVR machine and for the cross-validation grid, the `sklearn`[10] implementations were used. The cross-validated parameters uniquely corresponded to those of the learning machine, i.e. the regularization parameter, the kernel function, the polynomial degree, the bias parameter and the bandwidth (as they may apply).

For the Gaussian kernel

$$k_\gamma(x_i, x) = e^{-\gamma\|x_i - x\|_2^2} \tag{8}$$

[7] In order to compensate some possible skewing in feature distribution, we used the median statistic.

[8] A non-smoothed version of this comparison method for scalars is commonly used in acoustics for marginal sound level comparisons.

[9] http://radimrehurek.com/gensim

[10] http://scikit-learn.org

it was needed to estimate a reliable range of possible values for the bandwidth parameter $\gamma$. This parameter can become into a significant drawback in the case it is not carefully considered. In this sense we used the *median pairwise distance heuristic* (Gretton et al., 2012), which is denoted by $\widetilde{\gamma} \in \mathbb{R}_+$.

This statistical heuristic consists on computing all pairwise euclidean distances among training embeddings. The median of these distances is picked up as a reliable starting point for searching the value of the kernel bandwidth.

The SVR was trained over the 30 sentence embeddings inferred from Puces text. Up to 20 randomized searches were performed for each feature combination (section 7.1), where 20 machine estimators $f_{\alpha^*}$ for the $\mathcal{Y}$ candidature scores were obtained. The best 5 estimators were chosen for predicting candidature scores on RPM2 sentence embeddings.

Both for the regularization parameter $\lambda$ and for the kernel bandwidth $\gamma$ a random sampler was used. In the case of $\lambda$, the interval $[0.5, 2000]$ was randomly sampled. In the case of $\gamma$ in equation (8), the $\widetilde{\gamma}$ heuristic was used as mean value for searching over exponentially distributed random values. In Figure 1 a portrait of our proposed framework is showed.

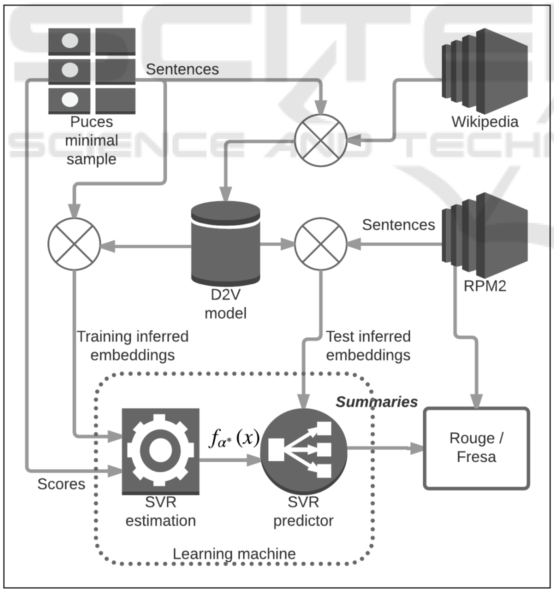

Figure 1: Portrait of our topic independent summarizer.

### 7.3 ATS Systems Evaluation

Learning a topic independent summarization model is for now our main aim. In order to assess the quality of the summaries obtained by means of such a learned model, a number of evaluation experiments were performed. By means of these evaluation experiments, we can know if needed features are actually present in a minimal sampling corpus.

Summary candidature score predictions made by $f_{\alpha^*}$ on RPM2 unseen sentence embeddings were evaluated. For this end, the Rouge and Fresa state-of-the-art ATS evaluation measures were used. In the case of Rouge evaluation, we used the mean of Rouge-$\{$1-2-3-SU4$\}$ measures.

In an effort to increase strictness of our study, we yielded 5% compression rate summaries (i.e. sentences with the highest predicted scores). Furthermore, they were truncated to the first 100 words prior to evaluation (which also applies to our baselines) (Nenkova, 2005).

In the final stage of our evaluation framework, we show some statistics which take into account some well known topic features, i.e. named entities, verbs, nouns, coreference mentions and even open domain Information Extraction triplets (openIE) (Fader et al., 2011; Li, 2015). Given these and other features that we propose that are underlying in sentence embeddings, we will show how important they are in human summaries, rather than what is the actual content they convey (this is why we consider them as language-use features). To this end, we selected our best 6 summaries as well as our 6 worst summaries[11].

The best and the worst summaries were indifferently chosen from both, the SVR and the Artex summarizers. After that, we used (7) to perform the logarithmic comparison between language-use features underlying in our best/worst summaries and those underlying in the human references. In order to compute the general human reference for a proposed set of language-use features, we randomly selected 6 human references from RPM2 about different topics (from four different humans). After that, the corresponding feature spectrum was plotted.

## 8 RESULTS

In this study, we used two well known performance measures for choosing $f_{\alpha^*}$ learned regression estimators, i.e. the weighted Pearson's correlation coefficient $\rho$ and the coefficient of determination $R^2$.

As part of our first results, we found that estimates provided by linear, polynomial and sigmoid kernels were completely unuseful for our proposed learning setting. This was because the low capacity these kernels induced to the learning machine, which produced high-subsampling estimates. Thus, we only

[11] Annotator statistics from (Nenkova et al., 2007) suggest 4-5 human references provide sufficient inter-annotator agreement stability. We used, however, 6 references.

show results for Gaussian kernels. In Figure 2 we show our best estimator for the word-LP-based sentence embeddings with $d = 10$ in (3). Notice that the imbalanced behavior of high scored samples is combined with the estimator's flattening on intermediate scores. This ultimately leads to subsampling, rather than leading to overfitting.

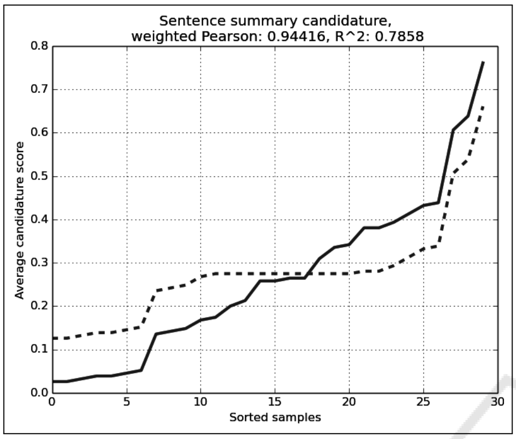

Figure 2: Regression estimation plot for the Puces word-LP-based sentence embeddings. The solid plot shows human references $\mathcal{Y}$ and the dashed plot shows the machine estimator $f_{\alpha^*}$.

From Table 1, it is observed that even although we have relatively high $\rho$ and $R^2$ coefficients between estimates and human references, there is no guarantee the output summaries to have good Rouge or Fresa measures (Table 2). Nonetheless, a singular fact is revealed: the highest performance measures were obtained for the worst feature combination, which is the PoS-based one. Analogously, the worst performance measures were obtained by our best combination, which is the word-LP-based one (columns 5 and 6 of Table 1; 2nd row of Table 2).

Other particular observations come from the values of $\lambda$ and $\gamma$. Notice that in general relatively small values of $\lambda$ performed well. This fact implies that relatively high learning regularization was needed, thereby acceptable generalization performance was reached. These observations are consistent with the fact that a minimal sample of features was learned from Puces. Among unreported unuseful results, we could observe the contrary scenarios.

Regarding to $\gamma$, it is observed the learning tends to select the bandwidth matching dominant features in the training embeddings. Even when we concatenated word-based embeddings onto PoS-based ones, the bandwidth approaches to those of word-based features. Notice in Table 1 the difference between bandwidths for PoS-based ($\gamma = 167.9$) and word-based ($\gamma = 71.50$) embeddings separately.

Table 1: Learning parameters for our most relevant (embedding) feature combinations.

| Embedding | $\widetilde{\gamma}$ | $\gamma$ | $\lambda$ | $\rho$ | $R^2$ |
|---|---|---|---|---|---|
| *Word-LP-based* | 0.004 | 10.61 | 1.00 | **0.944** | **0.785** |
| *Word-based* | 32.18 | 71.50 | 10.0 | 0.947 | 0.797 |
| *Word-PoS-based* | 14.17 | 71.31 | 5.80 | 0.948 | 0.791 |
| *PoS-based* | 29.59 | 167.90 | 5.00 | **0.955** | **0.814** |

Although there are uniform differences between $\widetilde{\gamma}$ and $\gamma$ for all combinations (2nd and 3rd columns of Table 1), the value of the difference for the case of word-LP-based embeddings is specially large (0.004 vs. 10.61, respectively). Even when word-based features are dominant, the two components added by sentence length $L$ and sentence position $P$ to word-based embeddings have relatively larger values than those that D2V derives. Thus, an equilibrium is observed: $\widetilde{\gamma}_{LP-based} \rightarrow 10.61 \leftarrow \gamma_{word-based}$.

Given that the Gaussian kernel resulted as the only useful one for our learning setting, it can be argued that language-use features are filtered from word-based embeddings by means of the SVR. For instance, in (Li et al., 2016; Liang et al., 2015) semantic features are easily filtered by using linear classifiers. This is also consistent with the fact the cosine measure (a linear function indeed) works quite well in semantic assessments. It clearly differs from *relevance detection*. As we will show later, this is what ultimately our method performed to produce summaries (Fader et al., 2011).

Regarding to sentence embeddings, we only reported results for $|c| = 16$ and $d = 10$ (3rd row, Table 2) because for the setting we are evaluating, using $d > 10$ (i.e. $d = 20, 25, ...; |c| = 8$) causes that the SVR to fall in overfitting. Our interpretation of this is that $d > 10$ gives redundant data, which results in noisy embeddings for our learning setting. Contrarily, for $d < 10$, summary candidature estimators showed high subsampling. This means that the SVR couldn't filter needed features from the embeddings.

There exist additional evidence that we think favors the language-use feature approach. It comes from prior empirical results, for instance in (Mikolov et al., 2013b; Gutman and Nam, 2015), well performance for $50 \leq d \leq 1000$ in semantic assessment was reported ($d \sim 300$ is the best trend). These parameter values are totally unuseful in our case. In general, for language-use features, other word-based embeddings different to ($d = 10, |c| = 16$) caused that the summaries to result in Rouge and Fresa measures even worse than those of our baselines (last rows, Table 2).

Another experiment of interest was performed by PoS-tagging the union between a subset of $10^5$ articles of the Wikipedia and the Puces document. A

D2V model was trained uniquely over the PoS tags of each article/sentence. After using only PoS-based embeddings for RPM2 candidature prediction, we found that those of $(d = 5, |c| = 16)$ had the best evaluation measure separately (6th row, Table 2). However, such measure was not better than the obtained via word-based sentence embeddings.

Table 2: Ranking of summary evaluation measures over the RPM2 dataset: the average of Rouge-{1-2-3-SU4} & Fresa.

| | Algorithm/embedding | **Rouge** | **Fresa** |
|---|---|---|---|
| 1 | *Artex* | **0.2021** | **0.0142** |
| 2 | *Word-LP-based* | *0.1867* | *0.0125* |
| 3 | *Word-based* | 0.1859 | 0.0120 |
| 4 | *Word-PoS-based* | 0.1824 | 0.0120 |
| 5 | *BL_rand* | 0.1796 | 0.0111 |
| 6 | *PoS-based* | 0.1767 | 0.0118 |
| 7 | *BL_1st* | 0.1754 | 0.0124 |
| 8 | *BL_long* | 0.1662 | 0.0116 |

We also explored the hypothesis of that, if we provide the combination between PoS-based and word-based sentence embeddings to the learning machine, the filtered compounding information could be richer (word-PoS-based, 4th row, Table 2). Therefore we concatenated the PoS-based onto the word-based sentence embeddings. These combined embeddings were fed to the learning machine for training. Resulting summaries were barely worse than those separately obtained via word-based sentence embeddings.

Once we observed the best evaluation measure for the word-based sentence embeddings so far, we decided adding two complementary dimensions to them: the sentence position $P$ and the corresponding sentence length $L$. Both these new components were added separately and together. We concatenated both $L$ and $P$ values together to the word-based embeddings. This feature combination became in the best result we reported in this work, i.e. the word-LP-based embeddings.

From Table 2, we observed in general the two best sentence embedding variants (the word-LP-based and the word-based) provide general features from Puces. This induces to our summarizer to be competitive to the Artex method. Herein the word-LP-based embeddings showed to be richer from the view of both kinds of evaluation measures we used. In this sense, word-based embeddings were not too distant. This fact is interesting given that the D2V embeddings are mainly designed to convey no other kind of features than co-occurrence ones.

As final part of our results, we report some statistic measurements. The finality of such measurements is to estimate the concentration of some simple features. Although we treat them as of underlying character, a pair of them show high concentration in human references. Thus the SVR is capable of filtering them from the embeddings, even when neither sentence position nor sentence length are present in word-based embeddings.

Our statistics were obtained from language-use features, which were detected by means of the coreNLP tool[12]. In Table 3 it is shown our proposed list of language-use features, which includes usual topic signatures like named entities, nouns and verbs. The cells show the concentration of each feature in six bad/good machine summaries, as well as in six summaries made by four different human annotators.

Given the feature concentrations of features in human references of Table 3, we computed the median reference $\varphi^R_{\mu_i}$. After that, we used Equation (7) to compute the feature spectrum $\mathcal{C}_{\mu_1}, ..., \mathcal{C}_{\mu_n}$ for Table 3. In Figure 3 we have represented the feature spectrum of bad summaries and good summaries, with respect to the reference $\varphi^R_{\mu_1}, ..., \varphi^R_{\mu_n}$ (which is a zero-valued plot, i.e. the inner semicircle in the plot).

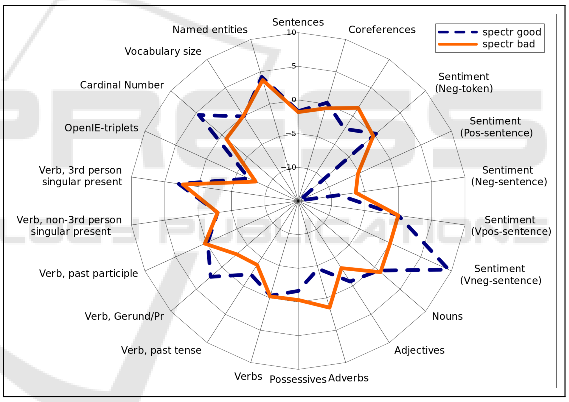

Figure 3: Feature spectrum of bad summaries and good summaries. The human references are represented by the zero-valued inner semicircle. The center of the circle is the most negative value.

According to Figure 3, there three main facts to be noticed. Firstly, there are features like named entities, nouns and all verb forms that are almost identical or nearly zero in concentration ratio, both in good and in bad summaries. Except for gerund verbs (e.g. *"it is growing"*), this means that the concentration of verbs is a non-informative feature (the same occurs for nouns). In the case of named entities, we have that both good and bad summaries showed values greater than zero. This means that machine summaries have an excessive concentration of named entities with respect to human references.

[12] http://stanfordnlp.github.io/CoreNLP/

Table 3: Language-use features obtained from the coreNLP tool. The features were extracted from six summaries about different topics. Six of them are good summaries and six are bad summaries. The remaining twenty four ones, were made by 4 different human references.

| **Language use feature** | **Good summs.** | **Bad summs.** | **Ref 1** | **Ref 2** | **Ref 3** | **Ref 4** |
|---|---|---|---|---|---|---|
| Sentences | 63 | 62 | 77 | 79 | 67 | 76 |
| Corefe-rences | 20 | 18 | 19 | 21 | 17 | 20 |
| Sentiment (Positive/ token) | 14 | 22 | 15 | 18 | 21 | 19 |
| Sentiment (Negative/ token) | 16 | 15 | 12 | 19 | 11 | 19 |
| Sentiment (Positive/ sentence) | 0 | 2 | 3 | 4 | 5 | 7 |
| Sentiment (Negative/ sentence) | 8 | 11 | 26 | 24 | 24 | 17 |
| Sentiment (VeryPos/ sentence) | 0 | 0 | 0 | 0 | 0 | 0 |
| Sentiment (VeryNeg/ sentence) | 2 | 0 | 0 | 0 | 0 | 1 |
| Nouns | 182 | 190 | 163 | 160 | 170 | 176 |
| Adjectives | 41 | 31 | 44 | 46 | 37 | 48 |
| Adverbs | 10 | 21 | 24 | 18 | 17 | 15 |
| Possessives | 11 | 13 | 15 | 18 | 11 | 12 |
| Verbs | 73 | 74 | 77 | 75 | 93 | 72 |
| Verb past tense | 16 | 13 | 20 | 22 | 19 | 21 |
| Verb Gerund Present | 8 | 4 | 7 | 4 | 7 | 5 |
| Verb past participle | 17 | 18 | 17 | 18 | 33 | 16 |
| Verb non-3rd person singular present | 4 | 4 | 8 | 3 | 8 | 4 |
| Verb 3rd person singular present | 13 | 12 | 9 | 6 | 9 | 11 |
| OpenIE-triplets | 59 | 52 | 134 | 109 | 132 | 145 |
| Cardinal Number | 25 | 13 | 15 | 15 | 14 | 9 |
| Vocabulary size | 329 | 333 | 335 | 331 | 333 | 329 |
| Named entities | 107 | 101 | 68 | 63 | 69 | 60 |

Secondly, some features are better in bad summaries with respect to references, i.e. the use of adverbs (e.g. *"we strongly recommend"*), the use of cardinal numbers and the presence of very negative sentiments by sentence.

Thirdly, except for very negative sentiments by sentence (the scale is 2:1, so probably it is not so confident), there are four very informative features: positive sentiments by sentence, negative sentiments by sentence and openIE triplets. Notice that the medians for positive/negative sentiments by sentence are in different scales (4.5 and 24, respectively), so the fact that machine summaries are lacking from both them is not contradictory. The case of openIE triplets is particularly interesting (e.g. factual phrases like *"John said something to Peter."*), because recently it was considered as a topic feature in the state of the art, rather than as a language-use feature (Ji et al., 2013; Li, 2015).

## 9 DISCUSSION

We estimate that the machine learns to filter the needed intersection of three context components: the language-use features, the length and position and the topics. The latter shows low importance in our case. In this framework a low-complexity and scale-invariant transformation $f_{\alpha^*}$ arises. It is thought due to the minimal text sample the learning machine needed to perform acceptably well, which occurs independently of the topics and the sizes of the training and test corpus.

The resulting analysis in last part of Section 8 provides sufficient arguments to determine the SVR used in our feature combination experiments could be improved. According to equation (5), all terms of the summation have the same bandwidth, which limits the learning machine either to a unique feature source (e.g. co-occurrence) or to a trade-off among 2 or more of them (e.g. co-occurrence, sentence length and sentence position). This limitation leads to subsampling when such feature sources are very distant in nature.

It is worth remarking that both Rouge and Fresa measures are based on the maximum lexical intersection hypothesis, which we have shown is not crucial to measure relevance. The bias in such a lexical hypothesis also manifests, on one hand, in high sensibility of the evaluation measures to the value of $L$. On the other hand, relatively high evaluation measures for some of our baselines take place.

Our statistical measurements showed some excessive, lacking and uninformative features of our machine summaries with respect to human references. Even when features like positive sentiments by sentence, negative sentiments by sentence and openIE triplets are very informative features, they are lacking in all our machine summaries (i.e. their concentration ratio in machine summaries is negative with respect to human references). Nonetheless, even when the scale of openIE triplets is relatively large (median reference equals to 133), their concentration ratio is not too negative. This leads us to think that their high concentration is detected in background by the SVR from D2V embeddings.

In the cases of positive/negative sentiments by sentence, we think that their importance is less impartial than in the case of openIE triplets. This is because we are analyzing a news corpus, so that the relevance of a sentence could be very biased by the literary genre, i.e. any human annotator is more likely to be overwhelmed by bad news than to be exited by good news. Furthermore, given the high negativeness of these two features, it is very possible that they were not detected by any of our summarizers. In this way, our approach provides insight about which features are needed to be reinforced and which ones are needed to be diminished.

Notice that the proposed approach in this work does not underestimates the use of topic features, but it only suggests that they are not dominant features for relevance detection. In topic-oriented ATS it is not different, because topic features can be seen as specific cases of language-use features.

## 10 CONCLUSIONS

We have defined a reasonable rigorousness in our evaluation framework. Therefore, according to our results, we have now an initial but sufficiently strong support for our proposed framework. The obtained support lead us to think that in order to maximize possibilities of keeping relevant contents from a document, it does not matter what is being said; but rather how it is being said.

In the short term, we propose to progressively increase the amount of training data. It could be observed that the Zipfian behavior of the human references clearly subsamples highly scored summary candidates. In this short term, we are also planning to explore the importance of a wider variety of state-of-the-art language-use features.

It is also considered to expand our experiments to more languages, to other well known datasets (e.g. DUC/TAC and MultiLing datasets) and to more sophisticated or task-driven learning machines. In this sense, a deep exploration of learned models is obligated in order to gain more comprehension of the task we are studying.

As we obtain better results, we propose to introduce ATS evaluation methods which are based on language-use features as main features for relevance detection. This in turn would motivate new ATS approaches seemed to the presented one in this work. In these methods relevance detection would be guided by a language-use feature spectrum followed by other more particular documents' aspects (as they are required, e.g. a given set of topics).

## ACKNOWLEDGEMENTS

This work is founded by The Mexican National Council for Science and Technology (CONACyT). Grant: 350326/178248. Thanks to The UNAM Computer Science graduate program and to The Laboratoire Informatique d'Avignon for their support.

## REFERENCES

Arroyo-Fernández, I. (2015). Learning kernels for semantic clustering: A deep approach. In *NAACL-HLT 2015 Student Research Workshop (SRW)*, pages 79–87.

Berg-Kirkpatrick, T., Gillick, D., and Klein, D. (2011). Jointly learning to extract and compress. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 481–490. Association for Computational Linguistics.

Cabrera-Diego, L. A., Torres-Moreno, J.-M., and Durette, B. (2016). Evaluating multiple summaries without human models: a first experiment with a trivergent model. In Métais, E., Meziane, F., Saraee, M., Sugumaran, V., and Vadera, S., editors, *Proceedings of the 21st International conference on the Application of Natural Language to Information Systems*, pages 91–101. NLDB.

Cao, Z., Wei, F., Dong, L., Li, S., and Zhou, M. (2015). Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159.

Cortes, C. (1995). *Prediction of Generalization Ability in Learning Machines*. PhD thesis, University of Rochester.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Divya, S. and Reghuraj, P. (2014). Eigenvector based approach for sentence ranking in news summarization. *IJCLNLP, April*.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Galanis, D., Lampouras, G., and Androutsopoulos, I. (2012). Extractive multi-document summarization with integer linear programming and support vector regression. In *COLING*, pages 911–926.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213.

Gutman, J. and Nam, R. (2015). Text classification of reddit posts. Technical report, New York University.

Harris, Z. S. (1968). *Mathematical Structures of Language*. Wiley, New York, NY, USA.

Hong, K. and Nenkova, A. (2014). Improving the estimation of word importance for news multi-document summarization. In *EACL*, pages 712–721.

Ji, H., Favre, B., Lin, W.-P., Gillick, D., Hakkani-Tur, D., and Grishman, R. (2013). Open-domain multi-document summarization via information extraction: Challenges and prospects. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 177–201. Springer.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73. ACM.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196.

Lee, J.-H., Park, S., Ahn, C.-M., and Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1):20–34.

Li, C., Liu, F., Weng, F., and Liu, Y. (2013). Document summarization via guided sentence compression. In *EMNLP*, pages 490–500.

Li, J., Li, J., Fu, X., Masud, M., and Huang, J. Z. (2016). Learning distributed word representation with multi-contextual mixed embedding. *Knowledge-Based Systems*.

Li, S., Ouyang, Y., Wang, W., and Sun, B. (2007). Multi-document summarization using support vector regression. In *Proceedings of DUC*.

Li, W. (2015). Abstractive multi-document summarization with semantic information extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1908–1913.

Liang, H., Fothergill, R., and Baldwin, T. (2015). Rosemerry: A baseline message-level sentiment classification system. *SemEval-2015*, page 551.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Moens, M.-F. and Szpakowicz, S., editors, *Workshop Text Summarization Branches Out (ACL'04)*, pages 74–81, Barcelona, Spain. ACL.

Louis, A. and Nenkova, A. (2009). Automatically evaluating content selection in summarization without human models. In *Conference on Empirical Methods in Natural Language Processing*, pages 306–314, Singapore. ACL.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

McDonald, R. (2007). *A study of global inference algorithms in multi-document summarization*. Springer.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Min, Z. L., Chew, Y. K., and Tan, L. (2012). Exploiting category-specific information for multi-document summarization. In *COLING*. ACL.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI*, volume 5, pages 1436–1441.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech Language Processing*, 4(2):1–23.

Nenkova, A. and Passonneau, R. J. (2004). Evaluating content selection in summarization: The pyramid method. In *Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*, pages 145–152, Boston, MA, USA.

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Bell System technical journal*, 3(2):324–346.

Saggion, H., Torres-Moreno, J.-M., da Cunha, I., and SanJuan, E. (2010). Multilingual summarization evaluation without human models. In *23rd International Conference on Computational Linguistics (COLING'10)*, pages 1059–1067, Beijing, China. ACL.

Sarkar, K. (2009). Sentence clustering-based summarization of multiple text documents. *International Journal of Computing Science and Communication Technologies*, 2(1):325–335.

Schölkopf, B., Bartlett, P., Smola, A., and Williamson, R. (1998). Support vector regression with automatic accuracy control. In *ICANN 98*, pages 111–116. Springer.

Shannon, C. E. (1949). Communication theory of secrecy systems*. *Bell system technical journal*, 28(4):656–715.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge UP. ISBN: 978-0-521-81397-6.

Torres-Moreno, J.-M. (2012). Artex is another text summarizer. *CoRR*, abs/1210.3312.

Torres-Moreno, J.-M. (2014). *Automatic text summarization*. John Wiley & Sons.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley New York.

Wan, X. and Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306. ACM.

Woodsend, K. and Lapata, M. (2012). Multiple aspect summarization using integer linear programming. In *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 233–243. Association for Computational Linguistics.