MACHINE LEARNING METHODS IN NATURAL LANGUAGE PROCESSING

by

Betül Güvenç

B.S., Mathematics, Bahçeşehir University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computational Science and Engineering

Boğaziçi University

2016

ACKNOWLEDGEMENTS

First of all, I want to express my thanks to my supervisor Assist. Prof. Fatih Ecevit for his time, guidance and endless support during the course of my thesis. I am also thankful to him for providing me an opportunity to explore myself.

With my deepest gratitude, I would like to thank Assoc. Prof. Atabey Kaygun for his continuous and never ending support during my study and research. He always kept me motivated with his unlimited patience when I lost my concentration, and guided me with his comprehensive domain knowledge when needed. I could not have imagined having a better mentor.

Besides my advisors, I am deeply grateful to Assist. Prof. Mohan Ravichandran for the long discussions that helped me sort out the technical details of my work, valuable comments and his endless support.

Last but not the least, I would especially like to express my gratitude to my family for their endless support, love and continuous support – both spiritually and materially throughout my life.

ABSTRACT

MACHINE LEARNING METHODS IN NATURAL LANGUAGE PROCESSING

There is a large number of algorithms for keyword extraction and text summarization in natural language processing, as we discuss some of these in this thesis. We started with a survey on automatic text summarization in order to understand the state of the art methods. Also we proposed a new and efficient method for keyword extraction task using Word2Vec and PageRank algorithms.

In this thesis, we investigated two different graph based text summarization algorithms for both single and multi-document settings on different types of texts where we used LexRank for multi-document summarization and TextRank for single document summarization. We also investigated a number of keyword extraction methods. Almost every keyword extraction method use high dimensional vectors to define words in a vector space. We approached the problem of automatic extraction of keywords from text as a unsupervised learning task and we treat each word in the document as a low dimensional vector. We developed a new keyword extraction method using Word2Vec and PageRank algorithms.

Our results show that summarization algorithms give best result on news texts, usable results on legal texts while they give less than optimal results for short stories. On the other hand, we also compared differences in using one-hot-representation and Word2Vec representation but we observed no significant differences between these methods.

ÖZET

DOĞAL DİL İŞLEMEDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Doğal dil işleme alanında çok sayıda anahtar kelime çıkarma ve metin özetleme algoritmaları vardır, bunlardan bazılarını bu tezde tartıştık. Methodları anlamak için otomatik metin özetleme üzerinde bir araştırma ile başladık. Ayrıca Word2Vec ve PageRank algoritmalarını kullanarak anahtar kelime çıkartmak için yeni ve etkili bir yöntem önerdik.

Bu tezde farklı metin tipleri üzerinde, hem tek metin hem de çoklu metin özetlemede kullanılan iki farklı grafik tabanlı metin özetleme algoritmasını araştırdık, çoklu metin özetlemede LexRank ve tekli metin özetlemede TextRank kullandık. Neredeyse tüm anahtar kelime çıkartma algoritmaları vektör uzayında kelimeleri tanımlamak için yüksek boyutlu vektörler kullanır. Biz metinden otomatik anahtar kelime çıkartma problemine öngörmesiz öğrenme işi olarak yaklaştık ve metindeki her kelimeyi düşük boyutlu vektör olarak ele aldık. Word2Vec ve PageRank algoritmalarını kullanarak yeni bir anahtar kelime çıkartma yöntemi geliştirdik.

Bizim sonuçlarımız gösteriyor ki özetleme algoritmalarımız haber metinleri üzerinde en iyi sonuç verirken kısa öyküler için daha az optimal sonuçlar vermektedir. Bunun yanında hukuki metinler üzerinde de kullanılabilir sonuçlar elde ettik. Öte yandan, one-hot temsili ve Word2Vec temsili kullanarak bu algoritmaların verdikleri sonuçların farklarını karşılaştırdık ama biz bu yöntemler arasında anlamlı bir farklılık gözlemleyemedik.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

h_k	Weighted sum of the neuron's input
idf	Inverse document frequency
p_{ij}	Transition probability
P	Transition probability matrix
R	The PageRank of page
tf	Term frequency
α	Dirichlet distribution parameter
β	Topics
η	The learning rate
λ	Eigenvalue
Φ	Multinomial parameter
π	Initial distribution
ρ	Corpus topic
$a(\eta)$	Log normalizer
$t(x)$	The sufficient statistic

LIST OF ACRONYMS/ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ART	Adaptive Resonance Theory
ATS	Automatic Text Summarization
CBOW	Continuous Bag of Words
EM	Expectation Maximization
GMM	Gaussian Mixture Model
IR	Information Retrieval
KEA	Keyphrase Extraction Algorithm
KL	Kullback Leibler
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MLE	Maximum Likelihood Estimation
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PAT	Patricia Tree
PDF	Probabilty Density Function
TF	Term Frequency
TF-IDF	Term Frequency Inverse Document Frequency
WWW	World Wide Web

1. INTRODUCTION

1.1. Natural Language Processing

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence (AI), and an area of computational science that investigates human languages. Basically it tries to model human language processes and generates useful information about human languages in a machine context. Because of the volume of the data and the inherent ambiguities in human languages, it is a challenging task. NLP provides algorithms for tasks such as machine translation, automatic summarization, keyword extraction, semantic search, question answering and information retrieval (IR) which are used by some search engines such as Google and Yahoo.

Developments on NLP go back to 40s. One of the first applications is done by Turing [1] during World War II. He proposed an intelligent system which does a machine translation that imitates a person in a conversation with a human. In the late 80s, machine learning approaches entered to NLP. One of the important studies in this direction is done by Berger *et al.* [2]. They created a maximum likelihood estimation to automatically construct maximum entropy models to use on several problems: bilingual sense ambiguition, word reordering and sentence segmentation. Most of the early studies used decision trees, however, later studies also used statistical based learning algorithms such as log linear models and Markov process models. The state of the art methods now can create more robust systems using statistical algorithms when system confronts with unfamiliar inputs and they can deal with large training corpora. Some systems use rules to make choices. These are called "rule based systems" and widely used in NLP. Unlike machine learning methods, they need to create more complex systems if they confront with unfamiliar inputs.

We divide NLP into two categories: natural language generation (NLG) and natural language understanding (NLU). Natural language generation constructs natural language from machine representation system with the result of the analysis of data.

On the other hand in natural language understanding algorithms try to understand the meaning of written natural language chunks then they create meaningful data.

It can be said that NLU is the exact opposite of NLG. While NLG is not obliged to manage ambiguities, NLU directly confronts with the challenge of understanding a text without ambiguity. NLG uses some level of linguistic representation of the text to create meaningful and grammatically correct linguistic units.

First NLG application was in machine tranlation systems. These applications analyze texts from input language then create a corresponding text in the target language. Report generation, document generation and mail merging are some of the application of NLG. Both categories have equally significant problems, nonetheless, there are less works in the literarure on NLG than NLU.

1.2. Automatic Text Summarization

A summary is defined to be "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that." [3]. Referring to this definition, summarizations can be produced from one or more documents, they should include important information about the texts and should be short. Accordingly, automatic text summarization (ATS) systems should consider these three details to create summaries from documents.

The history of the ATS goes back to the late 50s. However, the main impetus came in the 90s because of the exponential growth of the public usage of the World Wide Web (WWW). The increasing volume of information on the internet has exposed the need to quickly process textual data. ATS systems can reduce information overload, determine which documents worth reading and also provide a way to cluster similar documents and create summaries. However it is very difficult to implement these systems. One of the difficulties is evaluation of summaries coming out of text summarization systems, as there are no standard measures to evaluate these systems.

There are various classification systems on text summarization. The class of systems we are going to consider are distinguished on how sentences are created. We have two categories: extractive and abstractive. Extractive summaries select the most salient sentences of a document and use them in the summary without any modifications. They reduce the dimensionality of the vector representations of words by removing stop words and then stemming the words in the text. Abstractive summaries are deeper models than extractive ones. They analyse the source text on the semantic level in order to retrieve essential information from the text. After identifying the essential parts, they fuse these parts in a cohesive and grammatically correct way. In other words, there is a synthesis phase which involves natural language generation after analysis phase. Most of the studies in text summarization rely on extractive summarization systems. The main advantage is that it is easier to deal with extractive summaries than abstractive ones. However, extractive summaries may lack coherence and cohesion.

Text summarization can also be categorized as generic and query based. In generic summarization, it is assumed that the audience is a general one and topics in the documents have equal importance in creating summaries. On the other hand, query focused systems based on certain topics defined by a specific user query. In this approach, generating a useful summary needs to take a query and then tries to find information relevant to this specific user query.

Another important classification category for text summarization is the number of documents to be summarized. We have single document summarization and multi-document summarization. While single document summarization generates summaries from just one document, a multi-document summarizer creates a summary from a set of related source documents. A multi-document task is harder than single document summarization because of the redundancy problems and difficulty of achieving cohesion between the sentences generated from different documents. Some web based systems were inspired by research on multi-document summarizations such as Google News and Columbia NewsBlaster [4].

1.3. Keyword Extraction

Keyword extraction is related to automatic text summarization. While in text summarization we choose the most salient sentences to create summaries, in keyword extraction we choose the most salient words. In order to gain information from documents we need to identify which words are the most salient in the text. Basically, keyword extraction task can be described as an automatic identification of terms which best define the documents. The terms can be expressed as keyphrases, key terms or just keywords.

Keyword extraction task is a challenging area of natural language processing have a wide variety of applications. We now have large online document collections used by search engines and document databases. In large document corpora, keyword extraction which is also referred as "topic modelling" can be used to search, explore and analyse contents of documents. Due to the large size of the corpora, there is a strong demand for automatic keyword extraction systems.

There are two primary category for automatic keyword annotation: keyword assignment [5] and keyword extraction. In keyword assignment, keywords are selected from a controlled list of vocabulary which best describe document. In this method, only keywords in training data can be selected as a keywords for the new documents. For keyword extraction, keywords are chosen from the text instead of drawing from training data using various ranking methods.

1.4. Motivation

High quality keywords play crucial role to extract pertinent information from a given text. Moreover, with the rapid growth of online information it is difficult for human beings to accomplish this task. Keyword extraction is a challenging area of natural language processing in acceptable time because extracting most salient words from online texts is an expensive and time consuming task.

Scoring words based on text features and using machine learning methods to determine the feature weights have been studied for a long time in keyword extraction algorithms. There are various algorithms and systems that deal with keywords extraction tasks using such machine learning methods [6–9]. Most of the current studies use one-hot-representation to embed words in a vector space which mention in Section 4.1. However, in one-hot-representation a vector's dimensionalty is the same as the number of unique words in a text or corpus and this situation can cause to overfitting. Another major disadvantage of this approach is that this representation is computationaly expensive.

1.5. Contribution

In this thesis, we combined two different algorithms to create an efficient keyword extraction algorithm and investigated effects of this new method on computational complexity. We used Word2Vec algorithm to embed words in a low dimensional vector space. After that, we calculated similarity between words with cosine distance metric. In final step, to find ranking of each word we used PageRank algorithm. We also verified that this method is more efficient than the state-of-the-art methods in current use.

2. RELATED WORK

2.1. Automatic Text Summarization

The major challenge in automatic text summarization is distinguishing more informative parts of the text from the rest. Bulk of the work in this area rely on extractive methods. Earliest work in this area is done by Luhn *et al.* [10] where the authors used statistical information derived from word frequency and distribution to measure the importance of sentences on technical papers and magazine articles. Many ideas in this paper have formed the basis for automatic text summarization in later works. Baxendale *et al.* [11] found that finding salient parts of a text has also importance on summaries. Edmundson *et al.* [12] expanded Luhn's and Baxendale's ideas by combining them on multi-document summarizations. This work formed the foundation of the machine learning approaches in automated text summarization. In the 1990s as more features have become available more sophisticated techniques for deciding which sentences to extract are used by researchers as they integrated machine learning methods into summarization problems. However, many of these techniques do not take semantics into account. One of the first studies in this area is done by Kupiec *et al.* [13] where they use naive Bayes models to decide whether a particular sentence should be in the summarization. Later studies mostly used hidden Markov models [14] and log linear models [15] to find salient parts of texts. Apart from the machine learning methods, there were other approaches such as Barzilay and Elhaded *et al.* [16] where they used applied lexical chains in text summarization problems using WordNet [17] to identify these lexical chains.

The earliest study on multi-document summarization we could find is by McKeown and Radev in 1995. In this study, the authors summarized series of news articles in an abstractive way. However, their method was not suitable for extending to other domains because of its domain specific heuristics structure. In this direction Radev *et al.* [18] proposed another method to detect and use cluster centroids for generating summaries. Their work was an important contribution in natural language processing.

Because they introduced the bag of words model for the first time. Two significant similar methods are introduced in 2004 LexRank [19] and TextRank [20]. They are both inspired by cluster centroid method and PageRank [21] algorithm. TextRank does single document summarization and keyword extraction, LexRank is designed for multi-document summarization.

In recent years, generative hiererchical models have also been used for document summarization. For example Chang *et al.* [22] proposed sentence-based Latent Dirichlet Allocation (SLDA) model to select salient sentences. However, cost of the computation of this model was considereably high when it is compared with Latent Dirichlet Allocation and Latent Semantic Analysis [22].

Most of the studies we mention in this section on text summarization are done with English language. However, in the recent years the literature have been expanding on different languages such as Arabic, Chinese, Czech, English, French, Greek, Hebrew, Hindi, Romanian, Spanish [23]. While using summarization techniques in other languages, the most crucial problem is the challenging of generation of a manually annotated datasets for the summarization task.

2.2. Keyword Extraction

One of the most studied subjects of NLP is keyword extraction where we generate keywords to judge more quickly whether the text is worth reading or not. In a simple way, keywords define the main topics of texts or documents. Keyword extraction methods can be divided into four categories: simple statistics, linguistics, machine learning and other approaches.

Simple statistics methods are used frequently because we do not need training data and we focus on non-linguistic features of the text. Term frequency inverse document frequency (TF-IDF) [24] is the most commonly used algorithm. Some of the most known methods in this category which identify words statistically are n-gram statistics [25], word frequency, word co-occurrences [26] and PAT Tree (Patricia Tree) [27].

These methods are language independent, and therefore, the same technique can be used in different languages. The main disadvantage of such statistical methods is that sometimes the most important keyword may appear infrequently in the some specific texts such as health and medical records.

Linguistics methods use linguistics attributes of words, sentences or documents such as part-of-speech, syntactic structure and semantic qualities. Most common ones are lexical, semantic, syntactic and discourse analysis [28]. Lexical chain represents semantic information of a text in NLP. These approaches can be more accurate than statistical ones, however their computational cost are high and there is a need for language expertise.

Another significant keyword extraction approach is machine learning. Most keyword or keyphrase extraction methods are based on supervised learning algorithms. First supervised algorithm in this area is used by Turney *et al.* [6]. In this method the learning algorithm needs a set of features to learn and then using these features classifies words in the documents. TF-IDF and its variations, position of a phrase, POS information, and relative length of a phrase can be used as features in this approach. For example, the keyphrase extraction algorithm (KEA) is proposed by Frank *et al.* [7] uses a naive Bayes method to extract keyphrases from documents. Hulth *et al.* [8] show that using a part of speech tag as a feature lead to important improvement in the keyphrase extraction.

In supervised learning algorithms, there is a need for tagged document corpus which is difficult to build. For this reason unsupervised learning methods are used more frequently. One of the earliest unsupervised algorithm for keyword extraction is proposed by Muňoz *et al.* [9] in 1997 based on Adaptive Resonance Theory (ART). This is a neural network method to discover two-word keyphrases. This algorithm has the abilility of produce a large list of phrases but it has low precision. Graph based methods belong to the class of unsupervised algorithms. Mihalcea *et al.* [20] proposed the TextRank algorithm which depends on Google's PageRank algorithm to rank keywords based on the co-occurrence links between words.

3. BACKGROUND

3.1. Markov Chain Methods

3.1.1. Markov Chain

One of the most commonly used stochastic models in Natural Language Processing is Markov Chain and some models presented in this thesis take this form. In this section we present a general background on Markov chains.

A stochastic process is a family of random variables indexed by time which can either be discrete or continuous. In a stochastic process, random variable X_t depends on earlier values of the process. In the discrete case we write

$$P(X_t = j | X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) \tag{3.1}$$

We define a Markov process as a stochastic process which satisfies the Markov property which states that future states of the process depend on only present states, with a finite number of states and transitions between them. The difference between continuous and discrete Markov process is how time is treated. In the discrete case, time is a discrete variable and in continuous, time is a continuous variable holding values in the interval $[0, \infty)$.

A Markov chain is a discrete valued Markov process: Random variable X_t depends on only X_{t-1}. This means that given the present state X_{t-1}, future state X_t is conditionally independent of the past $(X_{t-2}, X_{t-3}, ...)$. This is also known as memoryless property. We have a set of states $X = (X_1, X_2, ...,)$ where X_i belongs to finite state space Ω and process starts from one initial state, and moves to another state. Each move is called as a step. When process is in state i, there is a probability p_{ij} which gives us the probability that the next state will be j. This probability does not depend on previous state probabilities. These probabilities are called transition probabilities.

$$P(X_t = j|X_0 = i_0, X_1 = i_1, ..., X_{t-1} = i_{t-1}) = P(X_t = j|X_{t-1} = i_{t-1}) = p_{ij} \quad (3.2)$$

The matrix P with elements p_{ij} is called the transition probability matrix of the Markov chain.

$$\begin{bmatrix} p_{00} & p_{01} & p_{02} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots \\ \vdots & \vdots & \vdots & \\ p_{i0} & p_{i1} & p_{i2} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

Figure 3.1. $n \times n$ transition probability matrix.

The distribution of a Markov chain is determined by its initial distribution and its transition matrix. We will call

$$\Pi = (\pi_{i1}, \pi_{i1}, ...) = (\pi_i^{(1)} \mid i \in \{1, 2, ..., n\}) \quad (3.3)$$

where each i is a state in the state-space, the initial distribution on X_n if

$$\pi_i^{(1)} = P(X_1 = i) \text{ and } \sum_{i=1}^{n} \pi_i^{(1)} = 1 \text{ for all } i \in \{1, 2, ..., n\}$$

$$P(X_{t+1}) = P(X_t)\pi^{(1)} \quad (3.4)$$

The n-step transition probability of a Markov chain is the probability that it goes from state i to state j in n transitions:

$$p_{ij}^n = P(X_{n+m}) \mid X_m = i) \quad (3.5)$$

The n-step transition matrix is

$$P^n = \{p_{ij}^{(n)}\} \quad (3.6)$$

With n-step transition matrix P^n, we can compute $(\pi^{(n)})$ with the relation

$$(\pi^{(n)})' = (\pi^{(1)})' P^{(n-1)} \tag{3.7}$$

Transition Probability Matrix Properties

A square matrix P is called

- positive, $P > 0$
  if $p_{ij} > 0$ for all $i, j \in \{1, 2, ..., n\}$
- non-negative, $P \geq 0$
  if $p_{ij} \geq 0$ for all $i, j \in \{1, 2, ..., n\}$

A nonnegative matrix P is called

- aperiodic, if for all $i \in \{1, 2, ..., n\}$
  $gcd\Big(k \in \mathbb{N} \mid (P^k)_{ii} > 0\Big) = 1$
  where *gcd* denotes the greatest common divisor
- irreducible, if for all $i, j \in \{1, 2, ..., n\}$ there exists a $k \in \mathbb{N}$ with $(P^k)_{ij} > 0$
- primitive, if and only if it is irreducible and aperiodic.
- stochastic, each row of P must be a probability vector, which requires that
  (i) $p_{ij} > 0$ for all $i, j \in \{1, 2, ..., n\}$
  (ii) $\sum_{j=1}^{n} p_{ij} = 1$ for all $i \in \{1, 2, ..., n\}$

We call π an invariant distribution of X , if its associated transition matrix satisfies

$$\pi' = \pi' P \tag{3.8}$$

and every Markov chain on a finite state space has at least one invariant distribution.

If $\pi^{(1)}$ equals an invariant distribution, then

$$\pi^{(k)} = \pi^{(1)} \quad \forall k \in \mathbb{N} \tag{3.9}$$

and we say that X is a stationary distribution.

Theorem 3.1. *Every eigenvalue λ of a Markov matrix satisfies $\mid \lambda \mid \leq 1$.*

Proof : Suppose $\lambda \in \mathbb{C}$ is an eigenvalue of A and $X \in V_n^{\mathbb{C}}$ is a corresponding eigenvector. Then

$$AX = \lambda X \tag{3.10}$$

Let k be such that $\mid x_j \mid \leq \mid x_k \mid$. For all j, $1 \leq j \leq n$. We mutiply each side of equation with k-th component

$$\sum_{j=1}^{n} a_{kj} x_j = \lambda x_k \tag{3.11}$$

Hence

$$\mid \lambda x_k \mid = \mid \lambda \mid \mid x_k \mid = \left| \sum_{j=1}^{n} a_{kj} x_j \right| \leq \sum_{j=1}^{n} a_{kj} \mid x_j \mid \tag{3.12}$$

$$\leq \sum_{j=1}^{n} a_{kj} \mid x_k \mid = \mid x_k \mid \tag{3.13}$$

Hence $\mid \lambda \mid \leq 1$. □

Theorem 3.2 (Perron-Frobenius)**.** *Any irreducible, aperiodic stochastic matrix P has an eigenvalue $\lambda = 1$ with unique associated left eigenvector.*

Proof : Suppose $AX = \lambda X$, X is a nontrivial eigenvector such that $X \in V_n^{\mathbb{C}}$, $X \neq 0$.

We use inequalities from Theorem 3.1 and reduce to

$$\mid x_k \mid = \left| \sum_{j=1}^{n} a_{kj} x_j \right| \leq \sum_{j=1}^{n} a_{kj} \mid x_j \mid \leq \sum_{j=1}^{n} a_{kj} \mid x_k \mid = \mid x_k \mid \tag{3.14}$$

This inequality and the Sandwich Theorem gives

$$\mid x_j \mid = \mid x_k \mid \quad for \quad 1 \leq j \leq n \tag{3.15}$$

Also, as equality holds in the triangle inequality section of inequalities (3.14), this forces all the complex numbers $a_{kj}x_j$ to lie in the same direction:

$$a_{kj}x_j = t_j a_{kk} x_k, \quad t_j > 0, \quad 1 \geq j \geq n \tag{3.16}$$

then we get

$$x_j = \tau_j x_k \tag{3.17}$$

where, $\tau_j = (t_j a_{kk})/a_{kj} > 0$. Then equation (3.15) implies, $\tau_j = 1$ and hence $x_j = x_k$ for $1 \leq j \leq n$. Finally,

$$\sum_{j=1}^{n} a_{kj} x_j = \lambda x_k = \sum_{j=1}^{n} a_{kj} x_k = x_k \tag{3.18}$$

So we get $\lambda = 1$. □

By the Perron-Frobenius Theorem, an irreducible and aperiodic Markov chain always converges to a unique stationary distribution.

Let $\pi > 0$ be a probability distribution over Ω. A Markov chain P is said to be reversible with respect to π if for all $x, y \in \Omega$ we have

$$\pi(x)P(x,y) = \pi(y)P(x,y) \tag{3.19}$$

A random walk on a $\mathbb{Z}$ is a special case of a Markov chain, $\mathbb{Z} = \{..., -1, 0, 1, 2, ...\}$ with $X_0 = 0$ and $P[X_{n+1} = X_n + 1] = p$, $P[X_{n+1} = X_n - 1] = 1 - p$.

A Markov chain can be represented by a directed graph with where each state is represented by a vertex and the transition probablity's by weighted edges. Random walk starts from some random vertex. Then given time step if we are in vertex x the next vertex y is selected according to the edge weight which is the transition probabilty p_{xy}. Random walks are used in the Pagerank algorithm where nodes are ranked on the using by their stationary probability.

3.1.2. Pagerank Algorithm

PageRank [21] is a widely known method proposed by Larry Page and Sergey Brin for counting citations of a website. It is used by Google internet search engine. Google uses the PageRank algorithm to determine the relevance or importance of a page. Google describes PageRank as

"PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites." [29]

If a page is linked to many pages, it has a high PageRank value, and if there are no links between other web pages there is no support for that page. PageRank can be calculated for collections of documents of any size. The PageRank value for a page u is defined recursively. It depends on the PageRank values of each page v linking to page u, divided by the number of links from page v.

$$R(u) = c \sum\nolimits_{v \in B_u} \frac{R(v)}{N_v} \tag{3.20}$$

$R(u)$ is the rank of web page u and c is a normalization factor. We let F_v be the set of forward links from u going to other pages and B_u be the set of links that come to u.

We define $N_v =\mid F_v \mid$ to be the number of forward links from u.

Let A be a square matrix where rows and columns are labeled by the web pages. A can be treated as an adjacency matrix which indicates whether or not there is a link between the pages. In the adjacency matrix, if there is an edge from node i to node j we let $adj_{ij} = 1$, if not we let $adj_{ij} = 0$. For weighted graphs, if there is an edge from node i to node j with weight w, then we let $adj_{ij} = w$ instead of using 1 to indicate a link. We let $A_{u,v} = \frac{1}{N_u}$ where N_u is the number of non-zero elements in each row. Then if R is treated as a vector over web pages

$$R = cAR \tag{3.21}$$

Equation (3.21) states that R is an eigenvector of A and $\dfrac{1}{c}$ is the eigenvalue of R. Since A satisfies the properties of a stochastic matrix, it can be treated as a Markov chain.

There is a limitation in this algorithm: If there are only two pages that link to each other, during the iteration process the algorithm will never converge because the it is trapped in a never-ending loop. To deal with this problem, random walk on graphs can be used. It is also called "random surfer"model. Page and Brin consider a surfer visits a random web page (a node of the web graph) with a certain probability which comes from the page's PageRank and executes a random walk on the web pages. At each step surfer goes to other pages that links to with equal probabilities. At the end of the process surfer visits some web pages more than the others. This means that more visited pages are more important according to PageRank algorithm. The main question in this algoritm is that what if the current location of the surfer has no out-links? In order to overcome this problem, a new factor is introduced which is damping factor [21], a scalar between 0 and 1. Page and Brin originally defined it as 0.85. It means that 85% of the time the surfer follows links at random, and 15% of the surfer goes to new link. So, the probability for the random surfer reaching one page is the sum of probabilities for the random surfer following links to this page.

The PageRank of page u is given as:

$$R(u) = \frac{1-d}{N} + d \sum_{v \in B_u} \frac{R(v)}{N_v} \tag{3.22}$$

where N is the total number of web pages, d is the damping factor and N_v is the number of forward links from web page v.

Equation (3.22) is an iterative algorithm that is guaranteed to terminate. The algorithm can be expresses as follows [21]:

$$\begin{aligned} R_0 &\leftarrow S \\ loop: & \\ R_{i+1} &\leftarrow AR_i \\ d &\leftarrow \| R_i \| - \| R_{i+1} \| \\ R_{i+1} &\leftarrow R_{i+1} + dE \\ \delta &\leftarrow \| R_{i+1} - R_i \| \\ \text{while } \delta > \varepsilon & \end{aligned}$$

Figure 3.2. PageRank algorithm.

S can be almost any vector over Web pages. Equation (3.22) can be also calculated algebraically as in Equation (3.20).

$$R = [dU + (1-d)A]R \tag{3.23}$$

where U is a square matrix with all elements equal to $1/N$, $[dU + (1-d)A]$ is the transition kernel which mixture of A and u transition kernel and it can also treated as a Markov chain. The PageRank value R is an eigenvector of this kernel.

3.2. Neural Network Methods

An Artificial Neural Network (ANN) is a classifier model which is inspired by biological systems such as the nervous system in the brain. ANN can detect relationship and patterns between inputs and outputs. They can cluster unlabeled data based on their similarities and also classify data when there are labeled training datasets. A simple single layer neural network is known as a perceptron.

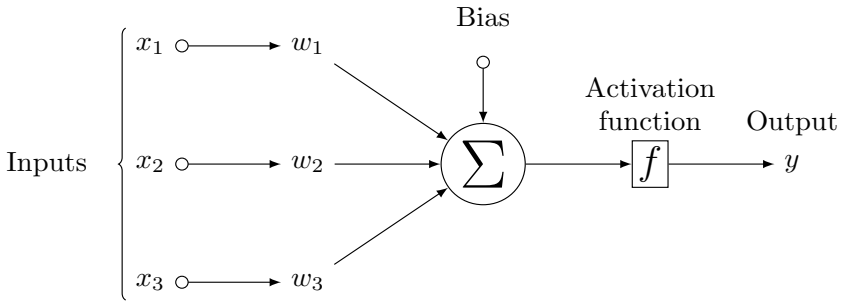

Figure 3.3. A single perceptron.

A perceptron finds a solution for splitting data linearly by iteratively learning the weights $(w_1, w_2, ..., w_m)$ and it classifies its inputs $(x_1, x_2, ..., x_m)$ into two categories. w_0 is known as bias unit which is always +1. In the simplest form, perceptron is defined by the equation:

$$net = \sum_{i=1}^{m} x_i w_i + w_0 \tag{3.24}$$

It takes a weighted sum of its inputs and this sum should be activated with an activation function.

$$y = f\left(\sum_{i=1}^{m} x_i w_i + w_0\right) \tag{3.25}$$

There are many activation functions that can be used in a perceptron. According to this function, if weighted sum is greater than some threshold, y will be 1 otherwise 0.

The most common used activation function is the sigmoid function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.26}$$

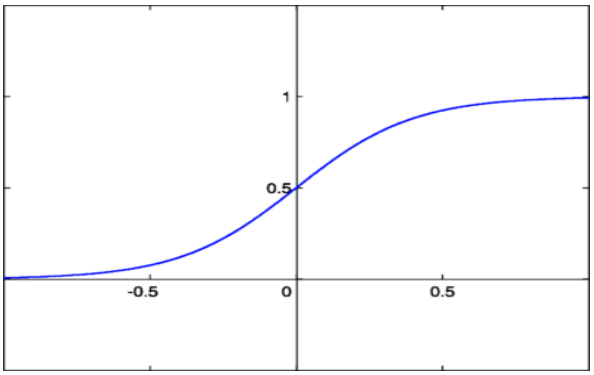

Figure 3.4. Sigmoid activation function.

This function never returns a 0 or a 1 because of its asymptopic nature.

During the training, weights are modified. The most common form of learning weights is adjusting them by using the difference between the target and the actual outputs. If classification is incorrect, weights are updated as follows

$$w_i = w_i + \eta x_i (y_i - o_i) \tag{3.27}$$

where η is the learning rate. When data are not linearly seperable then a multilayer perceptron is used. The simple architecture of a multilayer ANN consists of 3 layers: an input layer, a hidden layer and an output layer. If necessary, more then one hidden layer can be used between input and output layer. Each unit is fully connected to another units (Figure 3.5). Hidden units are the linear combination of input variables and weights, in the form of an sigmoid activation functions. The computation of a hidden unit for a simple multilayer neural network is given in Equation 3.28.

$$h_k = \sigma(w_k^T x + w_{k0}) = \frac{1}{1 + exp\left(-\sum_{j=1}^{d} w_{kj}^T x_j + w_{k0}\right)} \tag{3.28}$$

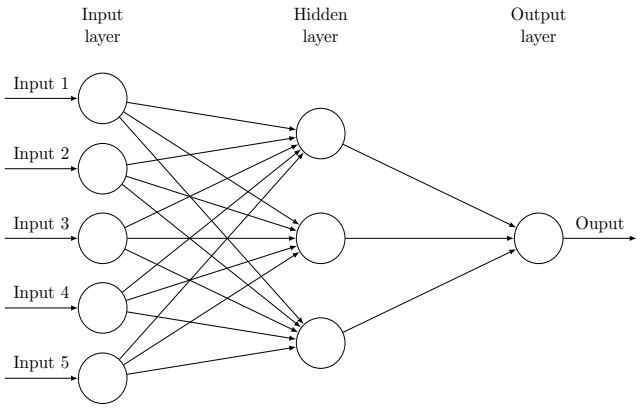

Figure 3.5. A multilayer perceptron.

In a multilayer ANN, the error and weight updates are computed with the back propagation algorithm. The goal of ANN algorithm is to train the network by adjusting the weights of each unit in order to minimize the error between the target output and the actual output. During the training process, algorithm starts with the initial weights and activations is propagated from input layer to hidden layer. Then the same is done for the hidden layer to the output layer. In the output layer the error function is calculated for each output neuron using the squared error function and we sum these errors over all neurons to get the total error and then we propagate the error back.

$$E_{total} = \sum \frac{1}{2}(y - o)^2 \tag{3.29}$$

The network updates its weights by taking the derivative of the error function until the error is acceptable. This update step is done by the delta rule.

The key idea behind the delta rule is to use gradient descent to search the hypothesis space of possible weight vectors that best fit the training examples [30].

$$\triangle W \propto -\frac{\partial E}{\partial W} \tag{3.30}$$

where

$$\triangle w_{kj} \propto -\frac{\partial E}{\partial w_{kj}} = -\frac{\partial E}{\partial o}\frac{\partial o}{\partial h_k}\frac{\partial h_k}{\partial w_{kj}} = -\sum_i (y_i - o_i)\frac{\partial o_i}{\partial w_{kj}} \tag{3.31}$$

This determines our update rule as

$$\triangle w_{kj} = \eta(y_k - o_k)\sigma^{'}(h_k)x_j \tag{3.32}$$

where h_k is the weighted sum of the neuron's input and x_j is the jth input. This error propagation is repeated by a weight updating for each node until the weights of the entire network are updated.

There are some parameters that need to be decided for a good optimization: optimal number of hidden layer, the learning rate, the number of repetitions which is also called epochs. These parameters affects the complexity of the network, learning time and ability to create accurate results. A smaller than optimal network architecture may not learn the problem and may cause high bias, while a larger than optimal network may cause in high variance. There is no common way to produce certain parameters. The most used technique to decide these parameters is state space search [31]. The algorithm starts with small networks and then constructs new networks by changing parameters.

3.2.1. Word2Vec

Word2vec is a shallow neural word embeddings model proposed by Mikolov *et al.* [32, 33] for vector representations of words. Here by "word embedding" we mean

model that maps each discrete word into a low-dimensional continuous vector-space from some raw text corpus. This model is an alternative to one-hot representation where the feature vector has the same length as the size of the vocabulary of the corpus which we mention in Section 4.1. However the latter model suffers from data sparsity.

Word2vec defines vector representations of words using relations between words using two different techniques: Continuous Bag-of-Words (CBOW) and Skipgram. Both of these methods describe how the neural network learns the word representations for each word. Since learning word representations is essentially an unsupervised algorithm, labels should be created for the given input to train the model depending on the architecture. Although there are other deep or recurrent neural network architectures generating word representations, Word2vec learns quickly relative to other models. The main problem with other methods is the relatively longer time required to train these models.

3.2.1.1. Continuous Bag of Words. Continuous bag of words model (CBOW) is trained to predict the target word with the contextual words that surround it, e.g. by taking the input words w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2} it predicts target word w_i. This model similar to the neural network language model, instead of using nonlinear hidden layer it uses linear projection layer which is shared for all words. It is called bag of words because order of words have no importance on the projection layer. If words are close together, this means that their meanings are somehow similar and the algorithm gives less weights to the distant words. During training, the error is backpropagated and since similar words appear in similar contexts, the vectors of similar words are updated towards similar directions so that they would predict the correct word. The CBOW model diagram is shown in Figure 3.2.1.1.

To create input layer, co-occurence matrix should be defined from the raw text. The input layer consists of the one-hot encoded input context $X = x_1, ..., x_C$. C is the number of context words which is also called the window size. It can be changed

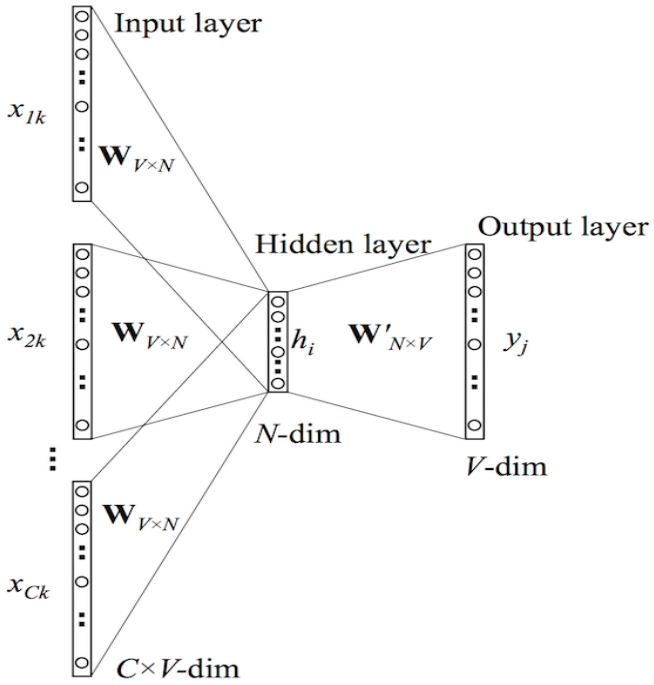

Figure 3.6. Continuous bag of word model (Mikolov et al., 2013a).

as a parameter and V is the cardinality of whole vocabulary. The hidden layer is an N-dimensional vector. The vectors of context words are connected to the hidden layer with a weight matrix W. The weight matrix is initialized to small random values as in any neural network before the training begins.

$$h = W^T X \tag{3.33}$$

Using Equation (3.33) we project vectors of context words to the hidden layer using weight matrix. Next the inputs of each node are computed in the output layer.

$$u = (W')^T h \tag{3.34}$$

$$u_j = (v')^T_{w_j} h \tag{3.35}$$

where v'_{w_j} is the j^{th} column of the output matrix W'. To calculate posterior probabilities of words, a multinomial distribution called the soft-max function can be used. Suppose $y = \varphi(u)$

$$y = p(w_t \mid w_1, ..., w_C) = \frac{exp(u)}{\sum exp(u')} \tag{3.36}$$

where y gives the probabilities of words. Here the word which has the highest probability is going to be our target word. To compute the error vector for the output layer, probability vector is subtracted from the target vector. Since the error is known, the weights are updated using backpropagation. According to Mikolov, this approach is faster than Skip-gram for large corpora [32].

3.2.1.2. Skipgram. The training objective of the Skip-gram model is to find word representations that are useful for predicting the surrounding words in a sentence or a document [33].

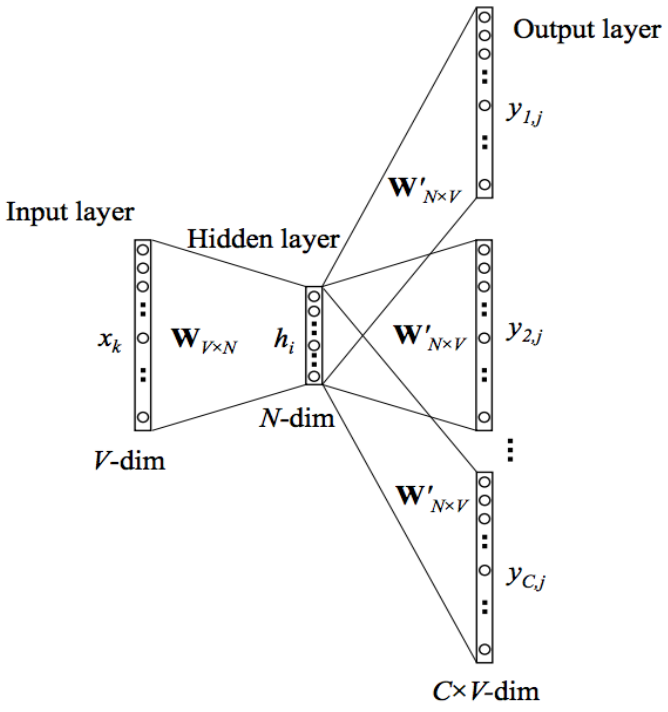

Figure 3.7. The Skip-gram model (Mikolov et al., 2013a).

Skip-gram model is the reverse of the CBOW model, where instead of using the context words in the input layer we use the target words. The hidden layer remains same and we find the context words in the output layer. This maximizes Equation (3.37).

$$p(w_{t-c}, w_{t-(c-1)}, ..., w_{t-1}, w_{t+1}, ..., w_{t+(c-1)}, w_{t+c}|w_t) \tag{3.37}$$

While one multinomial distribution is used to calculate output of CBOW model, Skip-gram uses C multinomial distribution in the output layer. Figure 3.7 shows the Skip-gram model.

According to Mikolov, Skip-gram is slower for infrequent words but works well with small training data [33].

3.2.1.3. Parameters. Representations of words can be adjusted by using CBOW and Skip-gram optimization algorithms. However, learning output vectors can be very expensive for large training corpora. To solve this problem, two different solutions are proposed in Mikolov *et al.* [33]. These are using hierarchical softmax and using negative sampling which we explain below.

The traditional softmax function computes conditional probabilities of all vocabulary words given the history. But this is computationally expensive especially on large corpora. Hierarchical softmax function is an efficient way of computing full softmax funtion. In the hierarchical log-bilinear model, the probability of the next word being w is the probability of making a sequence of binary decisions specified by how words are coded given the context [34]. The main advantage of hierarchical softmax function is that instead of evaluating W number of output nodes in the neural network to obtain the probability distribution, it is needed to evaluate only about $log_2 W$ number of nodes [33].

The negative sampling is an alternative optimization algorithm for the hierarchical softmax. In the negative sampling instead of updating the output vectors per iteration, we update only a sample of them. The target word vector should be in the sample and a probabilisty distribution should be arbitrarily chosen for sampling. This probabilisty distribution is called the noise distribution. Thus the task is to distinguish the target word w_O from draws from the noise distribution $P_n(w)$ using logistic regression, where there are k negative samples for each data sample [33]. It is shown that the Unigram distribution gives best result for the original paper.

3.3. Distributions

3.3.1. Dirichlet Distribution

Dirichlet distribution is a generalization of the Beta distibution over the $(K-1)$ dimensional simplex. It is also in the class of a Bayesian nonparametric models. It is parameterized by a parameter $\alpha = (\alpha_1, \alpha_2, ..., \alpha_K)$ where each $\alpha_i > 0$ which is a K dimensional vector. Let $\theta = (\theta_1, \theta_2, ..., \theta_K)$ and we say $\theta \sim Dirichlet(\alpha_1, \alpha_2, ..., \alpha_K)$ if

$$P(\theta_1, \theta_2, ..., \theta_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{3.38}$$

Dirichlet distribution is a member of exponential family distributions which has the following form:

$$p(x|\eta) = h(x) exp\{\eta^T t(x) - a\eta\} \tag{3.39}$$

where η is called natural parameter, $t(x)$ is the sufficient statistic, $h(x)$ is the underlying measure and $a(\eta)$ is log the normalizer which ensures that the density integrates to one.

$$a(\eta) = log \int h(x) exp\{\eta^T t(x)\} \tag{3.40}$$

The members of this family have important properties and have fundamental connections to the graphical models. The derivatives of the log normalizer gives the moments of the sufficient statistics and this gives an easy way to calculate expectation of the distribution in some intractable conditions. The expectation of the Dirichlet distribution is

$$E[(\theta_1, \theta_2, ..., \theta_K)] = \frac{(\alpha_1, \alpha_2, ..., \alpha_K)}{\sum_k \alpha_k} \tag{3.41}$$

When two different distribution are in the same family and also their parameters are part of the same parameter-space, it is said that these two distributions are a conjugate pair.

Theorem 3.3. *Dirichlet distributions is the conjugate prior of the Multinomial distribution.*

Proof: Given a sample $X = (x_1, ..., x_n)$ of multinomial data, $\theta \sim Dirichlet(\alpha_1, ..., \alpha_K)$ then $(\theta|X = x) \sim Dir(\alpha + x)$.

$$\begin{aligned} p(\theta|x_1, x_2, ..., x_n) &\propto p(x_1, x_2, ..., x_n|\theta)p(\theta) \\ &= \left(\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right) \left(\frac{n!}{m_1!...m_K!} \theta_1^{m_1}...\theta_K^{m_K}\right) \\ &\propto \frac{\Gamma(\sum_k \alpha_k + m_k)}{\prod_k \Gamma(\alpha_k + m_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k + m_k - 1} \\ &= Dir(\alpha_1 + m_1, ..., \alpha_K + m_K) \end{aligned} \tag{3.42}$$

□

where m_k defines the counts of instances of $x_n = k$. This situation shows that Dirichlet distribution is viewed as a distribution over parameters for the Multinomial distribution, where each sample from the Dirichlet distribution can be regarded as a Multinomial distribution.

3.3.2. KL - Divergence

Kullback-Leibler Divergence measures the difference between two probability distributions P and Q. For discrete distributions it is calculated as:

$$D_{KL}(P\|Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{3.43}$$

and for continuous distribituons it is calculated as :

$$D_{KL}(P\|Q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} \tag{3.44}$$

where p and d are the probability density functions (PDF) of the distributions P and Q. Kullback-Leibler distance is also described as expectation of $log(\frac{P}{Q})$ over the distribution Q.

$$D_{KL}(P\|Q) = E_P \left(log \frac{P}{Q} \right) \tag{3.45}$$

Note that KL - Divergence is not symmetric in P and Q.

3.3.3. Jensen's Inequality

Jensen's inequality is an important tool used in expectation maximization problems to find an adjustable lower bound on the log likelihoood. If there is a nonlinear relationship between input and output, Jensen's inequality states that for a convex function $f(x)$ we have

$$E(f(x)) \geq f(E[x]) \tag{3.46}$$

The reversed inequality is acceptable for a concave function

$$E(f(x)) \leq f(E[x]) \tag{3.47}$$

3.3.4. Expectation Maximization Algorithm

Expectation Maximization (EM) Algorithm is introduced by Dempster *et al.* [35] which is an iterative method to calculate maximum likelihood estimation of parameters in models which contain latent variables. There are two steps of EM to find the maximum likelihood estimation (MLE) of the marginal likelihood.

- Expectation step which is also called E step is used to compute the expected values of the log likelihood function with respect to the values of the observed data.
- Maximization step in which the observed and expected values of the observed and latent data are used to estimate the parameters of the model which will maximize the likelihood of the model given the data.

EM Algorithm: Iterate

1. **E-step: Compute** $q(x) = p(x|z;\theta)$
1. **M-step: Compute** $\theta = argmax_\theta \int_x q(x) log p(x,z;\theta) dx$

Figure 3.8. EM algorithm.

These two steps are repeated until the lower bound on the log likelihood converges. Suppose that x is a observed variable, z is latent variable and θ is the model parameter. EM solves MLE problem of the form

$$max_\theta log \int_x p(x,z;\theta) dx \tag{3.48}$$

$$q(x) = p(x|z;\theta) \propto p(x,z;\theta) \tag{3.49}$$

Expectation maximization algorithm can also be used for clustering. Gaussian mixture model (GMM) and K-means are some of the basic examples. GMM is also known as general version of K-means. K-means is a special case of hard EM where the covariances are spherical and the priors are equal. Here "hard" means that cluster memberships are not probabilistic.

Given the dataset $X = (x_1, ..., x_N)$ and the K number of clusters, the goal is to find a partition which minimize the objective fuction.

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \tag{3.50}$$

where S is the sum of partitions $S = \sum_{k=1}^{K} S_k$ and r_{nk} is equal to 1 if x_n belongs to cluster S_k, else $r_{nk} = 0$

K-Means Algorithm:

1. **Expectation Step:** Minimize J with respect to r_{nk}, μ_k is fixed.
2. **Maximization Step:** Minimize J with respect to μ_k, r_{nk} is fixed.

Repeat until convergence

Figure 3.9. K-Means algorithm.

4. MACHINE LEARNING FOR KEYWORD EXTRACTION

4.1. Keyword Extraction by Word2Vec

Various methods defining keywords for a given text or corpus have been used over the years. These methods are using different word representations in a vector space to feed algorithms with different attributes. There are different types of word representations in NLP. Most commonly used methods are distributional semantics and word embeddings. While distributional semantics defines vectors as a high dimensional vectors, word embedding models define them in a lower dimensional vector space.

The most common representation of distributional semantics is called one-hot representation in which dimensionalty is equal to vocabulary's cardinality. Elements of this vector space representation constist of 0's and 1's. However, this representation has some disadvantages. For example, in these representations it is difficult to make deductions about word similarity. Due to high dimensionalty they can also cause overfitting. Moreover, it is computationally expensive.

Word embeddings are designed to capture attributional similarities between vocabulary items. Words that appear in similar contexts should be close to each other in the projected vector space [36]. This means that grouping of words in a vector space must share same semantic properties. In word embeddings, Latent Semantic Analysis (LSA) [37] uses a count base dimensionalty reduction method. Word2Vec is created as an alternative [32, 33]. Its low dimensionalty can help to reduce computational complexity. Also compared with distributional semantics methods, it causes less overfitting. Word2Vec can also detect analogies between words [38].

Word2vec can make guesses about a word's meaning based on the past appearances. Those guesses can be used to cluster documents and classify them by topic.

Those clusters can be used in sentiment analysis recommendation systems and in keyword extraction methods. In this section we introduce a new method where we combine Word2Vec and PageRank algorithms for keyword extraction.

Our model takes Word2Vec representations of words in a vector space. Word2Vec is an unsupervised learning algorithm. The Word2Vec representations are constructed by neural network algorithm which is fed with a co-occurance matrix of vocabulary. Word2Vec has two different architecture to get word vectors: CBOW and Skip-gram. The original paper suggests using CBOW for large corpora with frequent words.

While we construct the Word2Vec model, we decide a threshold of counts of words because words that appear only once or twice in a large corpus are probably not interesting for the model, and there is not enough data to make any meaningful training on those words. A reasonable value for minumum counts changes between 0-100 and it depends on the size of corpora.

Another important parameter for Word2Vec model is the dimension of the vectors. This value changes between 100 and 400. Dimensions larger than 400 require more training but leads to more accurate models. We used Google News corpora which provided by Google which consist of 3 million word vectors [39]. We did not remove stop words or infrequent words because these algorithms use windows and to find vector representations. So we need the neighboring words to find vector representations.

Second step of this algorithm is to find PageRank value of each word. PageRank algorithm works with random walk. The original PageRank algorithm takes internet pages as a node. In our model PageRank algorithm takes Word2Vec representations of words. The cosine distance is used to calculate edge weights between nodes. TextRank algorithm uses a similar method. While TextRank chooses bag of word representations of words and a different similarity measure in finding edge weights, in this algorithm we used the Word2Vec representations and the cosine similarity. After PageRank values of words are found, we can get words which have the highest PageRank values. Finally these words can be seen as a keyword of a text.

4.2. Topic Modelling by Latent Dirichlet Allocation

Blei, Ng and Jordan presented the Latent Dirichlet Allocation (LDA) model and a Variational Expectation-Maximization (EM) algorithm for topic modeling. This is a generative probabilistic model that represents documents as a collection topics according to probabilistic representations of text and each topic is characterized by a distribution over words. LDA tries to find latent topics of given a set of documents. It is represented as a probabilistic graphical model (Figure 4.2). In this model, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corpus level parameters which are sampled once for each corpus while $\boldsymbol{\theta}$ is sampled once per document. $\mathbf{z}$ and $\mathbf{w}$ are the word level variables which sampled once for each word in each document.

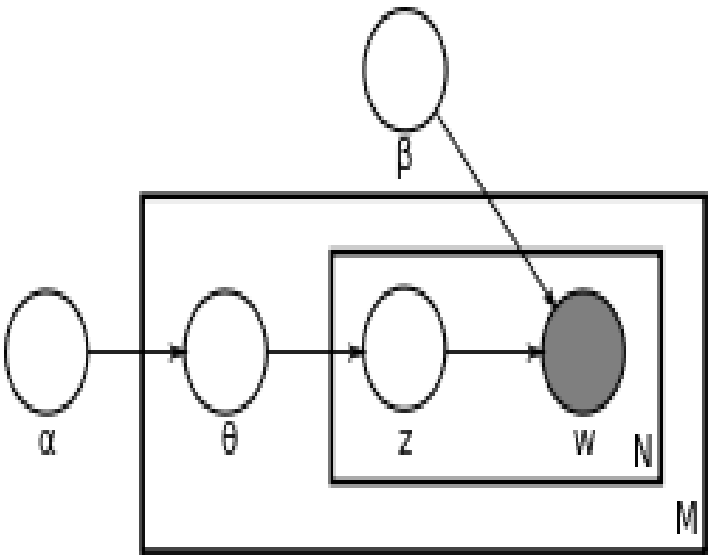

Figure 4.1. The LDA model (Blei et al., 2003).

In this model the only observed variables are words $\mathbf{w}$. The variable $\mathbf{z}$ is the hidden topic assignments and the aim of this model is to find topics. $\mathbf{M}$ is the number of documents in corpus and $\mathbf{N}$ is the number of words in each document. $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)$ represents topics. The variable $\mathbf{K}$ is known and is fixed in the algorithm. If we assume that α and β are known, this model can be written as follows:

$$p(w, z, \theta|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z|\theta)p(w|z, \beta) \tag{4.1}$$

It generates documents first the deciding number of words **N** the document will contain then it chooses $\boldsymbol{\theta}$ which determines the topic distribution for the document according to a Dirichlet distribution over a fixed set of **K** topics. In the next step, all words in the document are generated by first choosing topic assignment **z** according to multinomial distribution. Then we generate **N** words using the topic and topic assignments $p(w_n|z_n, \beta)$. Words are assigned to topics randomly and then we keep improving the model until model reaches an equilibrium. The posterior distribution of hidden variable is

$$p(z, \theta|w, \alpha, \beta) = \frac{p(w, z, \theta|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{4.2}$$

where $p(w|\alpha, \beta)$ is the marginal distibution of a document and we can express it as a summing over **z** and integrating over θ.

$$p(w|\alpha, \beta) = \int p(\theta, \alpha) \prod_{n=1}^{N} \sum_{k=1}^{K} p(z_n|\theta) p(w_n|z_{n,\beta_{1:K}}) \tag{4.3}$$

This equation is intractable to compute because of the coupling between α and β. It means that posterior distribution is intractable for exact inference. There are alternative ways to compute the posteriors using variational inference methods such as EM and Gibbs sampling.

The basic idea of variational inference is to use Jensen's inequality to find adjustable lower bound on the log likelihood. The variational distribution is created on the latent variables.

$$q(z, \theta|\gamma, \Phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\Phi_n) \tag{4.4}$$

In this equation γ is Dirichlet parameter and $\Phi = (\Phi_1, ..., \Phi_N)$ is the multinomial parameter, both known as free variational parameters. Next step is to set up an optimization problem to determine the values of Φ and γ . When optimization problem

is constructed, we use KL-Divergence to solve the problem.

$$(\Phi^*, \gamma^*) = argmin D_{KL}(q(\theta, z|\gamma, \Phi)||p(\theta, z|w, \beta, \alpha)) \tag{4.5}$$

Let us say $q(z, \theta|\gamma, \Phi)$ is q and $p(\theta, z|w, \beta, \alpha)$ is p then

$$\begin{aligned} D_{KL}(q||p) &= E_q[logq] - E_q[logp] \\ &= E_q[logq] - E_q[p(\theta, z, w|\beta, \alpha)] + log(w|\alpha, \beta) \end{aligned} \tag{4.6}$$

and using Jensen's inequality we bound $p(w|\alpha, \beta)$ as

$$\begin{aligned} logp(w|\alpha, \beta) &= log \int \sum_z p(\theta, z, w|\beta, \alpha) d\theta \\ &= log \int \sum_z \frac{p(\theta, z, w|\beta, \alpha) q(\theta, z)}{q(\theta, z)} d\theta \\ &\geqslant \int \sum_z q(\theta, z) log \frac{p(\theta, z, w|\beta, \alpha)}{q(\theta, z)} d\theta \\ &= E_q log \frac{p(\theta, z, w|\beta, \alpha)}{q(\theta, z)} \\ &= E_q[logp(\theta, z, w|\beta, \alpha)] - E_q[logq(\theta, z)] \end{aligned} \tag{4.7}$$

we denote it as $L(\gamma, \Phi; \alpha, \beta)$ then

$$D_{KL}(q||p) = -L(\gamma, \Phi; \alpha, \beta) + logp(w|\alpha, \beta) \tag{4.8}$$

$$logp(w|\alpha, \beta) = D_{KL}(q(\theta, z|\gamma, \Phi)||p(\theta, z|w, \beta, \alpha)) + L(\gamma, \Phi; \alpha, \beta) \tag{4.9}$$

If L is maximized with respect to γ and Φ, D_{KL} will be minimized.

$$\begin{aligned} L(\gamma, \Phi; \alpha, \beta) &= E_q[logp(\theta, z, w|\beta, \alpha)] - E_q[logq(\theta, z)] \\ &= E_q[logp(\theta|\alpha)] + E_q[logp(z|\theta|)] + E_q[logp(w|z, \beta)] \\ &- E_q[logq(\theta)] - E_q[logq(z)] \end{aligned} \tag{4.10}$$

When L is maximized depending on the equation with respect to Φ and γ, expectations of Φ and γ variables are found. There are two steps of EM algorithm. This step is called expectation step of the EM algorithm. In the maximization step, the lower bound is maximized with respect to α and β to find (marginal) log likelihood of the data. These two steps are repeated until the lower bound on the log likelihood converges.

Large corpus creates serious problems about sparsity. In this model a new document may have words that did not appear in the training corpus. Normally, maximum likelihood estimates assign zero probability to such words and new documents. The standard approach for coping with this problem is to "smooth" the multinomial parameters and assigning positive probability to all vocabulary items whether or not they are observed in the training set [40]. Blei, Ng and Jordan's proposed solution to this problem is to simply apply variational inference methods to the extended hierarchical model that includes Dirichlet smoothing on the multinomial parameter [41].

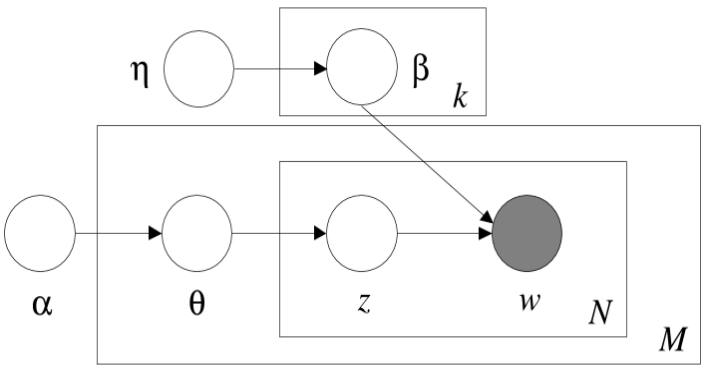

Figure 4.2. Smoothed LDA model (Blei et al., 2003).

This graphical model can be written as:

$$p(w, z, \theta|\alpha, \beta) = \prod_{i=1}^{k} p(\beta|\eta) \prod_{m=1}^{M} \left(p(\theta|\alpha) \prod_{n=1}^{N} p(z|\theta) p(w|z, \beta) \right) \tag{4.11}$$

Variational approach to Bayesian inference that places a separable distribution on the random variables β, θ, and z.

$$q(\beta_{1:K}, z_{1:M}, \theta_{1:M}|\rho, \Phi, \gamma) = \prod_{i=1}^{k} Dir(\beta_i|\rho_i) \prod_{d=1}^{M} q_d(\theta_d, z_d|\gamma_d, \Phi_{d,1:N}) \tag{4.12}$$

here ρ refers to corpus topics and γ document topics. This method is used for multi-documents and there is an additional update for the new variational parameter ρ. Main reason for adding new parameters is smoothing. Thus every word will be assigned with a to positive probabilty.

5. GRAPH BASED AUTOMATIC TEXT SUMMARIZATION

Graphs are widely used in natural language processing applications and they can readily encode the text in terms of their semantic and lexical structures. Graph-based summarization methods have been shown to be useful for both single document and multi-document summarization due to their ability to incorporate word frequency information into a formalized framework within which we can analyze sentence-to-sentence relationships easily [42]. These methods based on the assumption that most similar sections of the text are the most salient to the topic if each section is considered as a node. These sections can be words, sentences or paragraphs and edges represent the lexical or semantic connection between the two nodes. To calculate the significance of these nodes in a text, graph-based algorithms use ranking algorithms. The most common ones are PageRank [21] and HITS [43] algorithms. The selection of the most frequently visited nodes create summary of the input graph.

5.1. Representations of Documents

Unstructured documents should be transformed into a structured form in order to perform automatic text summarization with graph based methods by first transforming text using the bag-of-words representation. This model has been used for many years both information retrieval and text mining but it ignores grammar and word order. Bag of words representation represents document in a vector space. Sections of the text are treated as N-dimensional vectors where N is the number of words are counted in the bag and for each word a weight is calculated and corresponding to an entry in this vector. There are different ways of calculating these weight. Most popular one is TF (term frequency) in which we calculate the number of times each term occurs in the document. For multi-documents $TF - IDF$ (term frequency-inverse document frequency) is calculated for each term.

$$tf \cdot idf_t = tf_{t,d} \times log\left(\frac{D}{df_t}\right) \tag{5.1}$$

where D is the total number of documents in the collection and df_t is the number of documents in which term t appears, and $tf_{t,d}$ is the frequency of term t in document d.

There are some pre-processing steps are needed before documents we represented in a vector space.

Some words do not give any additional information while we select important parts of documents and they should be excluded from the vocabulary. These words are called stop words. There is a general way to determine stop words: Words are sorted and most frequent ones are selected. These words should be filtered by hand for their semantic content. In English, the stop words are words like 'the', 'and', 'which', 'on' [44].

Other important pre-process is stemming. Stemming is the process which reduces words to their roots. In this process we map related words to the same stem. This also helps to avoid high dimensionality.

For simplification, words are stemmed and stop words are removed from text. After we obtain a representation of document using the bag of words model in a vector space, the next step is creation of a fully connected graph to find similarities between nodes. Thus we obtain an adjacency matrix which is a 2-dimensional array of size $N \times N$ where N is the number of nodes in our graph. In case our graph is undirected this matrix is symmetric.

5.2. LexRank

LexRank is an extractive generic text summarization system proposed by Erkan *et al.* [19]. This algoritm is a stochastic, graph based method finding important sentences of any text in order to create meaningful summarizations in multi-document settings.

The main idea of this process is that relevant sentences in a cluster are more central and to find the most important sentences, one can use eigenvector centrality of the graph representation of sentences we construct above. There are two parameters we need to determine in this definition of centrality. First is the similarity between two sentences. Second is the computation of the overall centrality of a sentence given its similarity to other sentences [19]. There are different ways of defining similarity between two sentences, but in this discussion the cosine similarity metric is used.

The cosine of the angle between two vectors derived from their dot product:

$$x.y = \parallel x \parallel \parallel y \parallel \cdot cos(\Theta) \tag{5.2}$$

Given two vector of attributes, their cosine similarity is defined by using their Euclidean dot product

$$cos(x, y) = \frac{xy}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5.3}$$

Then the similarity between two sentences in multi-document is defined by the corresponding vectors:

$$idfmodified - cosine(x, y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2}\sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}} \tag{5.4}$$

where $tf_{w,s}$ is the number of occurrences of the word w in the sentence s and idf_w is the inverse document frequency of a word w.

Since we have well-defined representations of documents modeled as vectors (with TF-IDF counts) in a vector space, now we can define similarity between different documents in this space using their cosine similarity.

Erkan and Radev compared different centrality methods in their article. The LexRank algorithm which is based on the degree centrality is also proposed in the same article. In this method, every node represents one sentence and the graph is constructed using the cosine similarity between nodes. However, sentence pairs which their cosine similarity under a fixed threshold have not an edge to avoid complexity of calculation. This means, just significantly similar sentences are connected to each other. To find the centrality of sentences, number of edges are counted for each significant sentence which defines their degree. Then sentences with the highest degree are accepted as central sentences. The key parameter in this process is a proper cosine threshold. Very low and high thresholds may obscure many of the similarity relations in a text. The other important problem is computing the centrality of sentences over whole text. Ranking algorithms are frequently used for sentence extraction in natural language processing [45]. Sentences which have a high rank are more central to the topic.

In degree centrality, every edge has the same effect on the overall centrality. Every relationship in the graph is equally important. This could lead to some negative effects in the summary because it is possible that a collection of unimportant sentences may affect each other and increase the significance of each other. LexRank comes up with a new idea to solve this problem: Every node distributes its centrality to its neighbors. It uses both the PageRank algorithm and degree centrality. This idea can be expressed by the equation:

$$p(u) = \frac{1-d}{N} + d \sum_{v \in adj[u]} \frac{p(v)}{deg(v)} \tag{5.5}$$

where $p(u)$ is the centrality of node u, $adj[u]$ is the set of nodes that are adjacent to u, and $deg(v)$ is the degree of the node v.

This approach models the document as a graph like in the centrality degree, and then ranks each sentence to giving weights. The rank of each sentence is computed in an iterative manner. The algorithm re-computes the ranks of all sentences until a stopping condition is verified. The last step of this algorithm is choosing the sentences with the highest rank and in creating the summary. Some sentences can have similar

ranking values. This means that their meanings can be similar and taking both of them could cause to sentence multiplication in the summary. In order to prevent this problem, it can be used a cut off on ranking values of sentences. If one sentence' ranking value is equal to previous one or later one, it could be picked just random one of these for summary.

5.3. TextRank

TextRank is another graph based keyword extraction and text summarization method proposed by Mihalcea and Tarau [20]. TextRank applications rely on Google's PageRank algorithm to represent a text with a graph based ranking algorithm. First our text should be converted to graph. Vertices of our graph are going to be the terms of in the text and edges are going to represent the connections between two terms. The terms can be sentences,words or paragraphs of a text. Each edge is weighted indicating the importance of relationship between its nodes adjusted by the significance of the neighbours of these nodes like in the LexRank algorithm. For text summarization, sentences are chosen as vertices. Whereas LexRank is developed for multi-document summarization, TextRank is designed for single document applications. To create an edge, similarity measure defined as a function of content overlap is used. The overlap of two sentences can be determined simply as the number of common tokens between the lexical representations of the two sentences or it can be run through syntactic filters, which only count words of a certain syntactic category, e.g. all open class words, nouns and verbs, etc. [20].

$$Similarity(S_i, S_j) = \frac{\mid S_i \cap S_j \mid}{log(\mid S_i \mid) + log(\mid S_j \mid)} \tag{5.6}$$

S_i and S_j are sentences which are represented by the set of words. $S_i = w_1^i, w_2^i, ..., w_{n_i}^i$, $S_j = w_1^j, w_2^j, ..., w_{m_i}^j$. Number of common terms are divided by the length of the sum of two sentences to prevent the problem caused by very long sentences. In post-processing, weights of edges are calculated and then final ranking scores are calculated for each vertex. Sentences which have the highest rank are sorted to create a summary.

6. EXPERIMENTS AND RESULTS

6.1. Evaluation Methods

An important phase of automated text summarization and keyword extaction is evaluation of these tasks. However, evaluating results of these tasks is itself a difficult task, because the value of a summary or a keyword changes depending on a person, that is, evaluation methods are subjective. Sparck Jones states that "it is impossible to evaluate summaries properly without knowing what they are for" [46]. There are different evaluation methods for the comparison of summarization systems performance. These evaluation methods use system summaries and human generated summaries. Summarization evaluation methods can be classified into two categories: intrinsic and extrinsic.

Intrinsic evaluation methods measure how many main ideas of the source document are covered by the summary. An ideal summary is generated by human and it is compared with the system output. The quality of the summaries is measured by precision, recall and F-measure [47]. Both methods are used to evaluate keywords and summaries.

Extrinsic methods evaluate how helpful summaries are for a given task. While intrinsic methods measure quality of task, these mesure performence of task. Extrinsic methods usually require human effort, and therefore, more expensive than intrinsic ones. We prefered using extrinsic comparison methods in this thesis.

6.2. Text Summarization Experiments and Results

In this work we investigated two extractive text summarization methods which create text summaries by ranking and extracting sentences from the original documents: LexRank and TextRank. Both of these use PageRank algorithm. While LexRank generates multi-document summaries, TextRank generates single document summaries.

6.2.1. Experiments

In the following, we used LexRank algorithm to generate a multi-document summary for Proceedings of the Workshop on Automatic Text Summarization. For single document summaries we used TextRank algorithm to generate summaries for the three different text categories: short stories, legal texts and news articles [48–53]. For large texts, in generating cosine similarity matrix an entry is eliminated to avoid computational cost if the cosine value is lower than some threshold. However, we did not use any cutoff in our work because our datasets are not large. We extracted 5 sentences for each single document summarization experiment except for the short story summarization and our multi-document summarization.

For our experiments on multi-document summarization, we use articles from Proceedings of the Workshop on Automatic Text Summarization 2011 [54] in which there are five different articles about text summarization. Dataset contains approximately 19.000 words. After stemming and cleaning, there are 1600 unique words. Summary contains 10 salient sentences.

In our experiments on single text summarization, we first used the short story " The Nice People"by H.C. Bunner [50]. The second is the essay "Politics and the English Language" by George Orwell [51]. Each text is approximately 4000 words long and after stemming they contain approximately 900 unique words. We used TextRank to extract salient sentences. In our experiment, the summary of the short story contains 10 sentences, while the others contain 5 sentences. First, we picked 5 sentences for this text but the story contains a lot of dialogs and we could not get meaningful result. We increased the number of sentences in the summarization to 10 to get better results.

Our next experiment is done on two news articles. Most news articles are tightly orginized and dense, and therefore, suitable for such summarization algorithms. We took two news articles from Guardian newspaper [48, 49]. The first is a news article from their business section while the second one is an editorial on education system. They are both approximately 200 words long.

In our last experiment, we used two legal texts to be summarized. There have been development on creating more efficient text summarization algorithms for legal texts in the recent years [55,56]. Their content is different from other types of texts due to their statistics of words, probability of selection of textual units, relations between sentences, paragraphs and structures of the texts. One other important aspect of legal texts is that important information may appear only once in these texts. Because of these reasons, summarizing legal texts is a challenging subject and to get good results is difficult. The legal texts we used in our experiments consist of approximately 3500 words each [52,53]. After preprocessing, the documents hold approximately 500 unique words. First document is an opinion of the US supreme court about a case between Menominee Indian Tribe of Wisconsin and the US government [52]. Second text is a proposal for a visa regulation of the EU [53].

The results of text summarization experiments are included in Subsection 6.2.3. Main texts are given in the Appendix A.

6.2.2. Results

We obtained a mixed result from our multi-document summarization experiment. Since some sentences are about the specific algorithms, they broke the flow of the summary. The sentences 2 to 6 seem to be usable summary sentences. However, sentence 1 and the sentences 7 to 10 are not usable without proper context. The result indicates that even when the documents are higly related (such as a conference proceedings) if the source is not a single entity (a person or an organization) getting good results are difficult unless the documents show a high degree of homogeneity.

Our single text summarization experiments gave satisfactory results on the first news article as expected. As we mentioned above, the news articles are usually short and are usually about a single issue. In Figure 6.6, we picked the most salient sentences in the article (sentences 1 and 5). We also observe something interesting about the article: It prominently investigates the hedge fund Renaissance Technologies and its founder Simons (sentences 2,3 and 4).

The short story by Bunner did not yield a usable result in Figure 6.4. This is mostly due to the fact that the story contained a lot of dialogs and the pieces of these dialogs appeared in the summary without a proper context. In addition, there was another problem: Some sentences in the summarization is a part of two sentence sequence given in quotation. However, our summary picked only one of them. The article by Orwell also failed to produce a satisfactory result. Dense texts with many inter textual references such as Orwell's article are expected to produce poor results.

The legal texts, on the other hand, did produce usable results. In Figure 6.2, the summary correctly captured the fact that the text was about EU's new visa regulation for Kosovo residents (sentence 1), EU's recommendation on the issue (sentence 5) and its contextualization and reasons (sentences 2,3 and 4). However, there were some problems as well: sentence 3 in the summarization is a part of 2 sentence side-note given in brackets. But, our summary picked only the second one. In Figure 6.3, we again picked court's decision on the issue (sentence 3 and 5), its reasons and contexulazations (sentences 1, 2 and 4). However, sentence 2 again is a part of two sentence argument (indicated by the preceeding "Second,") in which we only picked one.

6.2.3. LexRank and TextRank Results

1. In future work, we will explore how we might take the graphic's intended message into account when identifying relevant paragraphs and will investigate the quality of extractive summaries of multimodal documents using our approach.
2. Automatic text summarization (ATS) is in many ways an encompassing subfield of NLP.

Figure 6.1. LexRank result on the proceedings of the workshop on automatic text summarization.

3. Researchers in the area often make use of part-of-speech (POS) tagging, named entity recognition (NER), language modeling, and many other techniques in NLP and machine learning.

4. Despite our plentiful access to these state-of-the-art tools and research, however, most complex ATS approaches rarely surpass the results achieved with simple statistics-based methods grown principally out of 60-year-old ideas of term frequency analysis [13, 7].

5. Nevertheless, more structured statistical approaches, based on Blei, et al.'s latent Dirichlet allocation (LDA) [3], have recently been showing promising results through the use of topic- or content-modeling [9, 8].

6. These approaches perform ATS by modeling input words as being generated from distinct hidden distributions of words.

7. In [9] and [8], salient words are seen as emanating from a different source than either background or document-specific words.

8. In this work, we present an even more highly structured statistical model where content words are modeled as being generated from a hierarchical topic structure where the most specific topics are at the bottom level and the most general topic forms the root.

9. Sentences are modeled as being made up of broad words that describe the input at a very general level, but also from more specific sub-topics that are arranged in a tree.

10. To build the tree, we make use of Bayesian nonparametric methods that allow the tree's structure to be organically generated directly from the input data.

Figure 6.1. LexRank result on the proceedings of the workshop on automatic text summarization (cont.).

1. Taking account of all the criteria which should be considered when determining on a case-by-case basis the third countries whose nationals are subject to, or exempt from, the visa requirement as laid down in Article -1 of Regulation (EC) No 539/2001 (as introduced by Regulation (EU) No 509/2014), the Commission has decided to present a legislative proposal to amend Regulation (EC) No 539/2001, transferring Kosovo from Annex I, Part 2 to Annex II, Part 4 of this Regulation.
2. Based on this assessment and given the outcome of the continuous monitoring and reporting that had been carried out since the launch of the visa liberalisation dialogue with Kosovo, the Commission confirms that Kosovo has met the requirements of its visa liberalisation roadmap on the understanding that by the day of the adoption of this proposal by the European Parliament and the Council, Kosovo will have ratified the border/boundary agreement with Montenegro and strengthened its track record in the fight against organised crime and corruption.
3. On the basis of this assessment and taking account of all the criteria listed in Article -1 of Regulation (EC) No 539/2001, it is appropriate to exempt persons from Kosovo from the visa requirement when travelling to the territory of the Member States.]
4. Regulation (EC) No 539/2001 was last amended by Regulation (EU) No 259/2014 8 when Moldova was transferred to the visa-free list after successfully implementing its Visa Liberalisation Action Plan; and by Regulation (EU) No 509/2014 9 when five Caribbean 10 and eleven Pacific countries 11 , as well as Colombia, Peru and the United Arab Emirates were exempted from the visa requirement – subject to the conclusion of visa waiver agreements between the EU and the respective third countries – following a periodical review of the visa lists.
5. The Commission committed to propose visa-free travel for persons from Kosovo for short stays (i.e. up to 90 days in any 180-day period) in the European Union once Kosovo had met all the requirements and other measures set out in the visa liberalisation roadmap.

Figure 6.2. TextRank result on the visa regulation of the EU.

1. After other tribal entities successfully litigated complaints against the Federal Government for failing to honor its obligation to pay contract support costs, the Menominee Tribe presented its own contract support claims to the IHS in accordance with the Contract Disputes Act of 1978 (CDA), which requires contractors to present each claim to a contracting officer for decision, 41 U. S. C. §7103(a)(1).
2. Second, the Tribe objects to the Court of Appeals' interpretation of the "extraordinary circumstances" prong as requiring a litigant seeking tolling to show an "external obstacl[e]" to timely filing, i.e., that "the circumstances that caused a litigant's delay must have been beyond its control."
3. The Court of Appeals denied the Tribe's request for equitable tolling by applying the test that we articulated in Holland v. Florida, 560 U. S. 631.
4. As the Tribe conceded below, see 614 F. 3d, at 526–527, it could not have been a member of the putative Cherokee Nation class because it did not present its claims to an IHS contracting officer before class certification was denied.
5. On remand, the District Court concluded that the Tribe's asserted reasons for failing to present its claims within the specified time "do not, individually or collectively, amount to an extraordinary circumstance" that could warrant equitable tolling.

Figure 6.3. TextRank result on the opinion of the US supreme court.

1. "I don't know how well he knows his own business, Major," I said as I started again for Brede's end of the veranda.

Figure 6.4. TextRank result on the short story "Nice People" by H. C. Bunner.

2. “I don’t want,” we heard Mr. Jacobus say, “to enter in no man’s privacy; but I do want to know who it may be, like, that I hev in my house.

3. “Oh, you poor, dear, silly children!” my wife cried, as Mrs. Brede sobbed on her shoulder, “why didn’t you tell us?”

4. “W-W-W-We didn’t want to be t-t-taken for a b-b-b-b-bridal couple,” sobbed Mrs. Brede; “and we didn’t dream what awful lies we’d have to tell, and all the aw-awful mixed-up-ness of it.

5. The Major (he was a widower) and Mr. Biggle and I looked at each other; and Mr. Jacobus, on the other side of the grape-trellis, looked at–I don’t know what–and was as silent as we were.

6. “I hain’t said I wanted to hev ye leave—-” began Mr. Jacobus; but Brede cut him short.

7. “But, my dear,”my wife said, gravely, “she doesn’t know whether they’ve had the measles or not.” “But, ” remonstrated Jacobus, “ef ye ain’t—-”

8. “Bring me your bill!” said Mr. Brede.

9. “Gentlemen,” said Mr. Brede, addressing Jacobus, Biggle, the Major and me, “there is a hostelry down the street where they sell honest New Jersey beer.”

10. But it seemed to us, when we looked at “our view,” as if we could only see those invisible villages of which Brede had told us–that other side of the ridges and rises of which we catch no glimpse from lofty hills or from the heights of human self-esteem.

Figure 6.4. TextRank result on the short story “Nice People” by H. C. Bunner (cont.).

1. Afterwards one can choose–not simply accept–the phrases that will best cover the meaning, and then switch round and decide what impressions one's words are likely to make on another person.
2. It is often easier to make up words of this kind (de-regionalize, impermissible, extramarital, non-fragmentary and so forth) than to think up the English words that will cover one's meaning.
3. Probably it is better to put off using words as long as possible and get one's meaning as clear as one can through pictures or sensations.
4. Nor does it even imply in every case preferring the Saxon word to the Latin one, though it does imply using the fewest and shortest words that will cover one's meaning.
5. As soon as certain topics are raised, the concrete melts into the abstract and no one seems able to think of turns of speech that are not hackneyed: prose consists less and less of words chosen for the sake of their meaning, and more and more of phrases tacked together like the sections of a prefabricated hen-house.

Figure 6.5. TextRank result on the essay "Politics and the English Language" by George Orwell.

Top 25 hedge fund managers earned $13bn in 2015 more than some nations
1. The world's top 25 hedge fund managers earned $13bn last year – more than the entire economies of Namibia, the Bahamas or Nicaragua.
2. Simons, a string theory expert and former cold war codebreaker, has made an estimated $15.5bn from Renaissance Technologies the mathematics-driven "quant" hedge fund he set up 34 years ago.
3. Despite the challenges, Simons and Griffin managed to increase their earnings by $500m and $400m, respectively, compared with last year.

Figure 6.6. TextRank result on the news article.

4. The fund, which is run from the tiny Long Island village of Setauket where Simons owns a huge beachfront compound, has donated $13m to Cruz's failed campaign.

5. The earnings of the best-performing hedge fund managers, published by Institutional Investor's Alpha magazine on Tuesday, dwarfs the pay of top Wall Street executives who have been under fire for their multimillion-dollar pay deals.

Figure 6.6. TextRank result on the news article (cont.).

Sats tests will harm next generation of writers, says Society of Authors

1. As year 6 children sit their Sats tests this week – including spelling, punctuation and grammar – the authors say that when the Department for Education introduces new terminology for grammatical structure, such as "fronted adverbs" and insists that exclamation marks can only end sentences starting with "what" or "how" it risks "alienating, confusing and demoralising children with restrictions on language just at the time when they need to be excited by the possibilities."

2. "Why do government ministers think they know more than teachers who have devoted their lives to the education of the nation's children?"

3. Author Anne Rooney, chair of the Society of Authors' educational writers group committee, attacked in particular the new rule on exclamation marks, saying that if children come across exclamation marks in books, they will wonder why the rules they have been taught don't match what they see in practice.

Figure 6.7. TextRank result on the editorial.

4. In a statement released by members of the Society of Authors who write for children and for education, they condemn current government policy on the teaching of writing and grammar. The Carnegie medal-winner David Almond, author of Skellig and a former teacher, added that children "instinctively know [that language] is a fluid, flexible, beautiful thing" and that they "learn how to talk, to sing, to converse by falling in love with language, by delighting in their own skills, by sharing and exploring those skills with others."
5. The statement calls on the government to "allow the current generation of schoolchildren in England to enjoy language, to be empowered by their skill in it, and not to become tangled in rules which have no application outside the narrow confines of a National Test."

Figure 6.7. TextRank result on the editorial (cont.).

6.3. Keyword Extraction Results

In this section we are going to use the terms of keyword extraction and topic modelling interchangeably. These aim to find the most important words in the documents. However, keyword extraction task is time consuming and computationally heavy process that requires lot of computing power, and there is no universal approach to define and compare different topics in a source. Also, topics in a text source are not always represented by nouns, they can also be represented by verbs or adjectives.

In our work, we used three different text datasets: legal texts, short stories and news articles [48–53]. These three are most frequently used categories in keyword extraction literature because there is a crucial need in the industry for classifiers which automatically categorize the data. We did not choose long texts because we used these texts for our text summarization experiments and long texts are not suitable in creating meaningful extractive summaries. Two different keyword extraction method were applied to these texts which we outlined in Section 4.1 and Section 4.2.

For this thesis and our Word2Vec experiments, word vectors trained on the Google News model provided by Google [39]. This model contains 300-dimensional vectors for 3 million words and phrases. In the pre-process we removed all stopwords and all words are converted lower case. After cleaning texts, we found vectors of all words from pre-trained text corpus and calculated similarities between words using the cosine distance metric. In the final step, PageRank algorithm was applied to the cosine similarity matrix and words which have highest PageRank value is taken as a topic of texts. The results are shown in Subsection 6.3.1.

In our LDA experiments, observed variables were words and we set unobserved variables as topics. Variational EM was used to find these topics. We explicitly fed the number of topics to the algorithm. We computed the optimum number of topics for each text by calculating the log likelihoods of the models for each topic number. The best number of topics is the one with the highest log likelihood value. We found that four is the best number. Although our texts are not long, we decided to choose ten topics instead of four. We have different types of texts, and since some of them are short, obtaining good keywords from these texts is challenging. If we increase the number of topics, we can reach more accurate keywords. However, some of these are going to be just noise. If we were used part of speech tagging in our algorithms, we could have eliminate these noisy keywords. The results are shown in Subsection 6.3.1.

6.3.1. Word2Vec and LDA Results

Word2Vec Results

dialogue	*visa*	*commission*	*undertaken*	*european*
february	*missions*	*dialogue*	*asylum*	*amending*

Figure 6.8. Keyword extraction results on the visa regulation of the EU.

LDA Results

behalf *refer* *act* *establish* *law*

import *passport* *may* *legisl* *four*

Figure 6.8. Keyword extraction results on the visa regulation of the EU (cont.).

Word2Vec Results

assistance *extraordinary* *contracts* *nearly* *errors*

revision *seq* *indian* *treated* *beyond*

LDA Results

tribal *outsid* *author* *pace* *case*

order *caus* *order* *within* *classact*

Figure 6.9. Keyword extraction results on the opinion of the US supreme court.

Word2Vec Results

knew *puck* *two* *skin* *ridges*

thought *talk* *combings* *path* *sure*

LDA Results

leav *account* *dear* *better* *side*

bite *rather* *began* *dear* *across*

Figure 6.10. Keyword extraction results on the short story "Nice People" by H. C. Bunner.

Word2Vec Results

fraternity *passages* *five* *heard* *visual*

lacking *english* *wherever* *prove* *strong*

LDA Results

averag *conscious* *drop* *desir* *root*

mani *mass* *result* *involv* *idiom*

Figure 6.11. Keyword extraction results on the essay "Politics and the English Language" by George Orwell.

Word2Vec Results

m *donated* *biggest* *dropped* *year*

strated *republicans* *according* *equivalent* *alpha*

LDA Results

pay *best* *back* *pay* *accord*

said *announc* *clinton* *campaign* *renaiss*

Figure 6.12. Keyword extraction results on the news article.

Word2Vec Results

children *encourage* *know* *children* *children*

test *confines* *fluid* *author* *novelists*

LDA Results

mark *test* *amongst* *say* *risk*

societi *sat* *statement* *skill* *mark*

Figure 6.13. Keyword extraction results on the editorial.

7. CONCLUSION

In this work, we did a survey on two extractive text summarization methods which create text summaries by ranking and extracting sentences from the original documents. Also as a new method proposed an approach to create an efficient keyword extraction method using Word2Vec and PageRank algorithms.

We approach the problem of automatic extraction of keywords from text as a unsupervised learning task and we treat each word in the document as a low dimensional vector. We used Word2Vec algorithm to embed words in a low dimensional vector space. After that, to find ranking of each word vector we used PageRank algorithm. In final step we calculated similarity between words with cosine distance metric. After PageRank values of words are found we get the keywords from the highest PageRank values.

In our topic extraction experiments in which we used LDA we used two different representation: one-hot-representation and Word2Vec. But we observed no significant differences between results obtained from LDA using one-hot-representation and Word2Vec representation.

We also investigated graph based text sumarization methods: LexRank and TextRank. Both method use PageRank algorithm to rank sentences. After we obtain a representation of documents using the bag-of-words model in a vector space, the next step is creation of fully a connected graph to find similarities between nodes. In both method, every node represents one sentence and the graph is constructed using the different similarity measures between nodes. This approach models the documents as a graph and then ranks each sentence to giving weights. Both algorithms are stochastic, graph based method finding important sentences of a texts in order to create meaningful summarizations in multi-document or single document settings.

We used three different text categories for both application: short stories, legal texts and news articles [48–53]. Our results show that summarization algorithms give best result on news articles and the legal texts but, they give less than optimal results for short stories 6.2.3.

7.1. Future Work

As a future work, Word2Vec model can be used for summarization algorithms. After calculating each word vector of a sentence, its word vectors can be added and this summation vector represents this sentence. As we did in keyword extraction, important sentences can be found using PageRank graph algorithm. Nodes of PageRank would consist of sentence vectors and using cosine similarity one can measure weigth between sentences. Sentences which have highest PageRank value then would be taken as a salient sentences.

Furthermore, our text summarization model may be developed by using semantic information embedded in the text. There are some noise words in our sentences which we choose as a salient such as 'second', 'he'or 'she'. While we look summarization as a whole these words break the flow of the summary. For instance, when a sentence start with 'second', we should find the sentence starting with 'first'within a window to form a meaningful summary.

As we see in the results, in some experiments LDA gave good results and other experiments gave usable results with Word2Vec algorithm. One can use the LDA results together with Word2Vec results to get better keywords.

7.2. Accomplishments

In this thesis, we developed a new keyword extraction method combining two different algorithms: Word2Vec and PageRank. We investigated effects of this new method. Almost every keyword extraction method use high dimensional vectors to define words in a vector space. We approach the problem of automatic keyword ex-

traction from text as a unsupervised learning task and we used Word2Vec algorithm to embed words in a low dimensional vector space. After PageRank values of words are found, we can take words which have highest PageRank value as keywords. We also verified that this method is more efficient than the state-of-the-art methods in current use.

REFERENCES

1. Turing, A. M., "Computing machinery and intelligence", *Mind*, Vol. 59, No. 236, pp. 433–460, 1950.

2. Berger, A. L., V. J. D. Pietra and S. A. D. Pietra, "A maximum entropy approach to natural language processing", *Computational linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

3. Radev, D. R., E. Hovy and K. McKeown, "Introduction to the special issue on summarization", *Computational linguistics*, Vol. 28, No. 4, pp. 399–408, 2002.

4. McKeown, K. R., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman and S. Sigelman, "Tracking and summarizing news on a daily basis with Columbia's Newsblaster", *Proceedings of the second international conference on Human Language Technology Research*, pp. 280–285, Morgan Kaufmann Publishers Inc., 2002.

5. Witten, I. H., G. W. Paynter, E. Frank, C. Gutwin and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction", *Proceedings of the fourth ACM conference on Digital libraries*, pp. 254–255, ACM, 1999.

6. Turney, P. D., "Learning algorithms for keyphrase extraction", *Information Retrieval*, Vol. 2, No. 4, pp. 303–336, 2000.

7. Frank, E., G. W. Paynter, I. H. Witten, C. Gutwin and C. G. Nevill-Manning, "Domain-specific keyphrase extraction", *IJCAI*, Vol. 99, pp. 668–673, 1999.

8. Hulth, A., "Improved automatic keyword extraction given more linguistic knowledge", *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 216–223, Association for Computational Linguistics, 2003.

9. Munoz, A., "Compound key word generation from document databases using a hierarchical clustering ART model", *Intelligent Data Analysis*, Vol. 1, No. 1, pp. 25–48, 1997.

10. Luhn, H. P., "The automatic creation of literature abstracts", *IBM Journal of research and development*, Vol. 2, No. 2, pp. 159–165, 1958.

11. Baxendale, P. B., "Machine-made index for technical literature: an experiment", *IBM Journal of Research and Development*, Vol. 2, No. 4, pp. 354–361, 1958.

12. Edmundson, H. P., "New methods in automatic extracting", *Journal of the ACM (JACM)*, Vol. 16, No. 2, pp. 264–285, 1969.

13. Kupiec, J., J. Pedersen and F. Chen, "A trainable document summarizer", *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 68–73, ACM, 1995.

14. Conroy, J. M. and D. P. O'leary, "Text summarization via hidden markov models", *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 406–407, ACM, 2001.

15. Osborne, M., "Using maximum entropy for sentence extraction", *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*, pp. 1–8, Association for Computational Linguistics, 2002.

16. Barzilay, R. and M. Elhadad, "Using lexical chains for text summarization", *Advances in automatic text summarization*, pp. 111–121, 1999.

17. Miller, G. A., "WordNet: a lexical database for English", *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

18. Radev, D. R., H. Jing and M. Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user stud-

ies", *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pp. 21–30, Association for Computational Linguistics, 2000.

19. Erkan, G. and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligence Research*, pp. 457–479, 2004.

20. Mihalcea, R. and P. Tarau, "TextRank: Bringing order into texts", Association for Computational Linguistics, 2004.

21. Page, L., S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", *Stanford InfoLab*, Citeseer, 1999.

22. Chang, Y.-L. and J.-T. Chien, "Latent Dirichlet learning for document summarization", *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 1689–1692, IEEE, 2009.

23. Giannakopoulos, G., "Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop", *Proceedings of the MultiLing 2013 Workshop on Multilingual Multidocument Summarization*, pp. 20–28, 2013.

24. Sparck Jones, K., "A statistical interpretation of term specificity and its application in retrieval", *Journal of documentation*, Vol. 28, No. 1, pp. 11–21, 1972.

25. Luhn, H. P., "A statistical approach to mechanized encoding and searching of literary information", *IBM Journal of research and development*, Vol. 1, No. 4, pp. 309–317, 1957.

26. Matsuo, Y. and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information", *International Journal on Artificial Intelligence Tools*, Vol. 13, No. 01, pp. 157–169, 2004.

27. Chien, L.-F., "PAT-tree-based keyword extraction for Chinese information re-

trieval", *ACM SIGIR Forum*, Vol. 31, pp. 50–58, ACM, 1997.

28. Ercan, G. and I. Cicekli, "Using lexical chains for keyword extraction", *Information Processing & Management*, Vol. 43, No. 6, pp. 1705–1714, 2007.

29. *Facts about Google and Competition*, 2011, `http://www.google.com/competition/howgooglesearchworks.html`, accessed at June 2016.

30. Mitchell, T. M. *et al.*, *Machine learning. WCB*, Vol. 8, McGraw-Hill Boston, MA:, 1997.

31. Kwok, T.-Y. and D.-Y. Yeung, "Constructive algorithms for structure learning in feedforward neural networks for regression problems", *Neural Networks, IEEE Transactions on*, Vol. 8, No. 3, pp. 630–645, 1997.

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

33. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality", *Advances in neural information processing systems*, pp. 3111–3119, 2013.

34. Mnih, A. and G. E. Hinton, "A scalable hierarchical distributed language model", *Advances in neural information processing systems*, pp. 1081–1088, 2009.

35. Dempster, A. P., N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

36. Turney, P. D., "Similarity of semantic relations", *Computational Linguistics*, Vol. 32, No. 3, pp. 379–416, 2006.

37. Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic analysis", *Journal of the American society for infor-*

mation science, Vol. 41, No. 6, p. 391, 1990.

38. Mikolov, T., W.-t. Yih and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations.", *HLT-NAACL*, pp. 746–751, 2013.

39. Mikolov, T., W.-t. Yih and G. Zweig, *word2vec*, 2013, `https://code.google.com/p/word2vec/`, accessed at June 2016.

40. Jelinek, F., *Statistical methods for speech recognition*, MIT press, 1997.

41. Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

42. Nenkova, A. and K. McKeown, "Automatic summarization", *Foundations and Trends in Information Retrieval*, Vol. 5, p. pp. 103–233, 2011.

43. Kleinberg, J. M., "Authoritative sources in a hyperlinked environment", *Journal of the ACM (JACM)*, Vol. 46, No. 5, pp. 604–632, 1999.

44. Fox, C., "A stop list for general text", *ACM SIGIR Forum*, Vol. 24, pp. 19–21, ACM, 1989.

45. Mihalcea, R., "Graph-based ranking algorithms for sentence extraction, applied to text summarization", *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, p. 20, Association for Computational Linguistics, 2004.

46. Jones, K. S. *et al.*, "Automatic summarizing: factors and directions", *Advances in automatic text summarization*, pp. 1–12, 1999.

47. Powers, D. M. W., "Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation", *School of Informatics and Engineering Technical Reports (24p ver.of ECAI'2008 Evaluation Evaluation)*, , No. SIE-07-001, 2007.

48. Neate, R., *Top 25 hedge fund managers earned 13bn dollar in 2015 – more than some nations*, 2016, https://www.theguardian.com/business/2016/may/10/hedge-fund-managers-salaries-billions-kenneth-griffin-james-simon, accessed at June 2016.

49. Flood, A., *Sats tests will harm next generation of writers, says Society of Authors*, 2016, https://www.theguardian.com/books/2016/may/11/sats-tests-will-harm-next-generation-of-writers-says-society-of-authors, accessed at June 2016.

50. Bunner, H. C., *The Nice People*, http://www.gutenberg.org/ebooks/10947?msg=welcome_stranger, accessed at June 2016.

51. Orwell, G., *Politics and the English Language*, http://gutenberg.net.au/ebooks02/0200151.txt, accessed at June 2016.

52. Menoninee Indian Tribe of Wisconsin v. United States, 577 U. S. 510 (2016).

53. Council of European Union, *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulation (EC) No 539/2001*, 2016, http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2016:0277:FIN, accessed at June 2016.

54. *Proceedings of the Workshop on Automatic Text Summarization*, Collocated with Canadian Conference on Artificial Intelligence, St. John's, Newfoundland and Labrador, Canada, 2011.

55. Moens, M.-F., "Summarizing court decisions", *Information processing & management*, Vol. 43, No. 6, pp. 1748–1764, 2007.

56. Farzindar, A. and G. Lapalme, "Legal text summarization by exploration of the thematic structures and argumentative roles", *Text Summarization Branches Out Workshop held in conjunction with ACL*, pp. 27–34, 2004.

APPENDIX A: SOURCE TEXTS

A.1. VISA REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulation (EC) No 539/2001 listing the third countries whose nationals must be in possession of visas when crossing the external borders and those whose nationals are exempt from that requirement

(Kosovo*)

*This designation is without prejudice to positions on status, and is in line with UNSCR 1244/1999 and the ICJ Opinion on the Kosovo declaration of independence.

EXPLANATORY MEMORANDUM

1. CONTEXT OF THE PROPOSAL

- **Reasons for and objectives of the proposal**

The European Commission launched a visa liberalisation dialogue with Kosovo on 19 January 2012. On 14 June 2012, it presented to Kosovo a roadmap, which identified all the legislation and other measures that Kosovo needed to adopt and implement to advance towards visa liberalisation. The Commission committed to propose visa-free travel for persons from Kosovo for short stays (i.e. up to 90 days in any 180-day period) in the European Union once Kosovo had met all the requirements and other measures set out in the visa liberalisation roadmap.

Figure A.1. Visa Regulation of the European Parliament and of the Council.

The Commission insisted on sufficient progress in readmission and reintegration as necessary elements to be put in place before launching a visa liberalisation dialogue with Kosovo.With a set of important reforms implemented since 2011, Kosovo made satisfactory progress in establishing a functional policy framework for the reintegration of returnees in Kosovo, as it had already done in the case of readmission. The Commission continued to monitor and assess, in its regular reports, Kosovo's progress in enhancing its readmission framework and the effective reintegration of returnees.

The visa liberalisation roadmap contained two sections: Section I addressed readmission and reintegration; Section II, four separate 'blocks' of the visa dialogue. The four blocks of the visa roadmap comprised specified requirements in document security; border/boundary and migration management, including asylum; public order and security; and fundamental rights related to the freedom of movement. Kosovo was first requested to adopt or amend in line with the EU acquis the legislation set out in the roadmap and then fully implement it.

The Commission conducted the visa dialogue with Kosovo in reinforced consultation with the Council, notably by involving the Council in developing the visa roadmap and with the full participation of Member States' experts in assessing Kosovo's progress in fulfilling the requirements of the roadmap.

The visa dialogue with Kosovo has been conducted without prejudice to Member States' position on status. [1]

The European Union Rule of Law Mission in Kosovo (EULEX KOSOVO), in line with its mandate, [2] has played an important role in monitoring, mentoring and advising Kosovo on adopting and implementing the reforms and fulfilling the requirements set out in the roadmap. Effective cooperation by Kosovo with EULEX,

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

including in discharging its executive mandate, has been essential.

Since launching the visa dialogue, the Commission has presented regular reports to the European Parliament and to the Council on its assessment of Kosovo's fulfilment of the requirements of the roadmap. These reports addressed requirements related both to readmission and reintegration and the different blocks of the visa roadmap. Each report drew upon information provided by Kosovo; assessment missions undertaken by the Commission and Member States' experts to assess Kosovo's progress in the different blocks of the visa dialogue and data provided by EUROPOL, FRONTEX, EASO and EULEX.

The Commission has adopted until now three reports on Kosovo's progress in the visa dialogue — the first one on 8 February 2013, [3] the second on 24 July 2014: [4] the third on 18 December 2015 [5] , complemented by the fourth one adopted today. [6] These reports contained an assessment of progress by Kosovo in fulfilling the requirements of the visa roadmap, recommendations addressed to Kosovo and an assessment of the potential migratory and security impacts of visa liberalisation.

In its third report, the Commission set out eight recommendations corresponding to eight outstanding requirements of the visa roadmap, including four key priorities. It noted the border/boundary delineation agreement with Montenegro should be ratified by Kosovo before visa free status is granted to persons from Kosovo.

In its report accompanying the present proposal, the Commission observed that Kosovo had taken important steps towards fulfilling the requirement of ratifying its border/boundary agreement with Montenegro and fulfilled sufficient elements of building up its track record in the fight against organised crime and corruption.

Based on this assessment and given the outcome of the continuous monitoring and

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

reporting that had been carried out since the launch of the visa liberalisation dialogue with Kosovo, the Commission confirms that Kosovo has met the requirements of its visa liberalisation roadmap on the understanding that by the day of the adoption of this proposal by the European Parliament and the Council, Kosovo will have ratified the border/boundary agreement with Montenegro and strengthened its track record in the fight against organised crime and corruption.

Taking account of all the criteria which should be considered when determining on a case-by-case basis the third countries whose nationals are subject to, or exempt from, the visa requirement as laid down in Article -1 of Regulation (EC) No 539/2001 (as introduced by Regulation (EU) No 509/2014), the Commission has decided to present a legislative proposal to amend Regulation (EC) No 539/2001, transferring Kosovo from Annex I, Part 2 to Annex II, Part 4 of this Regulation.

As indicated in the roadmap, this amendment only covers the individuals from Kosovo who are holders of a biometric passport issued in compliance with International Civil Aviation Organisation (ICAO) standards and EU standards for security features and biometrics in travel documents [7].

- **Consistency with existing policy provisions in the policy area**

Council Regulation (EC) No 539/2001 lists the third countries whose nationals must be in possession of a visa when crossing the external borders of the Member States and those whose nationals are exempt from that requirement. Regulation (EC) No 539/2001 is applied by all Member States – with the exception of Ireland and the United Kingdom – and also by Iceland, Liechtenstein, Norway and Switzerland. The Regulation is part of the EU's common visa policy for short stays of 90 days in any 180-day period.

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

Kosovo is currently listed in Annex I, Part 2 of Regulation (EC) No 539/2001, i.e. among those entities and territorial authorities that are not recognised as states by at least one Member State. Persons from those entities are required to hold a visa when travelling to the territory of EU Member States.

Regulation (EC) No 539/2001 was last amended by Regulation (EU) No 259/2014 [8] when Moldova was transferred to the visa-free list after successfully implementing its Visa Liberalisation Action Plan; and by Regulation (EU) No 509/2014 [9] when five Caribbean [10] and eleven Pacific countries [11] , as well as Colombia, Peru and the United Arab Emirates were exempted from the visa requirement – subject to the conclusion of visa waiver agreements between the EU and the respective third countries – following a periodical review of the visa lists. On 9 March 2016 and 20 April 2016, the Commission made proposals to amend Regulation (EC) No 539/2001, transferring – respectively - Georgia [12] and Ukraine [13] to the visa-free list.

The criteria which should be taken into account when determining – based on a case-by-case assessment – the third countries whose nationals are subject to, or exempt from, the visa requirement are laid down in Article -1 of Regulation (EC) No 539/2001. They include "illegal immigration, public policy and security, economic benefit, in particular in terms of tourism and foreign trade, and the Union's external relations with the relevant third countries, including in particular, considerations of human rights and fundamental freedoms, as well as the implications of regional coherence and reciprocity"[14]. Particular attention should be paid to the security of travel documents issued by the third countries concerned.

Kosovo has already exempted all EU citizens from the visa requirement for stays of up to 90 days within 6 months. Should this decision be revoked or should the visa-free regime be abused, the reciprocity and suspension mechanisms of Regulation

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

(EC) No 539/2001, as amended by Regulation xxx can be activated.

- **Consistency with other Union policies**

On 6 April 2016, the Commission proposed setting up an EU Entry/Exit System (EES) to strengthen the Schengen area's external borders [15] . The main objectives of this proposal are to improve the quality of border checks for third country nationals and to ensure a systematic and reliable identification of overstayers. The future EES will thus be an important element to ensure lawful use of the visa-free stays in the Schengen area by third country nationals and to contribute to preventing irregular migration of nationals from visa-free countries.

Furthermore, in its Communication of 6 April 2016 [16] , the Commission announced that it will assess the need, feasibility and proportionality of the establishment of an EU Travel Information and Authorisation System (ETIAS). The Commission has committed to explore still in 2016 whether such an alternative layer of control for visa-free nationals is feasible and proportional, and will effectively contribute to maintaining and strengthening the security of the Schengen area.

2. LEGAL BASIS, SUBSIDIARITY AND PROPORTIONALITY

- **Legal basis**

As the proposal will amend the EU's common visa policy, the legal basis for the proposal is point (a) of Article 77(2) of the Treaty on the Functioning of the European Union (TFEU). The proposed regulation will constitute a development of the Schengen acquis.

- **Subsidiarity, proportionality and choice of the instrument**

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

As Regulation (EC) No 539/2001 is a legal act of the EU, it can only be amended by way of an equivalent legal act. Member States cannot act individually to achieve the policy objective. No other (non-legislative) options to achieve the policy objective are available.

3. RESULTS OF EX-POST EVALUATIONS, STAKEHOLDER CONSULTATIONS AND IMPACT ASSESSMENTS

- **Stakeholder consultations**

Regular discussions with Member States in the Council Working Party on the Western Balkans (COWEB), as well as regular exchanges with the European Parliament on the visa liberalisation process have taken place.

- **Collection and use of expertise**

The Commission has collected comprehensive data on Kosovo's implementation of all requirements of the visa liberalisation roadmap. The Commission's fourth report is accompanied by a Commission staff working document setting out the potential migratory and security impacts of visa liberalisation for Kosovo, as well as the set of measures that Kosovo has implemented since December 2015 to prevent an irregular migration crisis. [17]

- **Impact assessment**

In the above staff working document, the Commission provided an updated analysis and statistical information on the possible migratory and security impacts of visa liberalisation for persons from Kosovo, as well as the set of measures that

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

Kosovo has implemented since December 2015 to prevent an irregular migration crisis, based on input provided by relevant EU agencies and other stakeholders. No further impact assessment is necessary.

- **Detailed explanation of the specific provisions of the proposal**

Regulation (EC) No 539/2001 will be amended, transferring Kosovo from Annex I, Part 2 (visa-required list) to Annex II, Part 4 (visa-free list). A footnote will be added specifying that the visa exemption will be limited to holders of biometric passports issued in line with the standards of International Civil Aviation Organisation (ICAO) and EU standards for security features and biometrics in travel documents (Council Regulation (EC) No 2252/2004).

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

amending Regulation (EC) No 539/2001 listing the third countries whose nationals must be in possession of visas when crossing the external borders and those whose nationals are exempt from that requirement

(Kosovo*)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular point (a) of Article 77(2) thereof,

Having regard to the proposal from the European Commission,

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

After transmission of the draft legislative act to the national parliaments,

Acting in accordance with the ordinary legislative procedure,

Whereas:

(1) Council Regulation (EC) No 539/2001 [19] lists the third countries whose nationals must be in possession of a visa when crossing the external borders of the Member States and those whose nationals are exempt from that requirement. The composition of the lists of third countries in Annexes I and II should be, and should remain, consistent with the criteria set out therein. References to third countries in respect of which the situation has changed as regards those criteria should be transferred from one annex to the other, as appropriate.

(2) The criteria which should be taken into account when determining – based on a case-by-case assessment – the third countries whose nationals are subject to, or exempt from, the visa requirement are laid down in Article -1 of Regulation (EC) No 539/2001. They include "illegal immigration, public policy and security, economic benefit, in particular in terms of tourism and foreign trade, and the Union's external relations with the relevant third countries, including in particular, considerations of human rights and fundamental freedoms, as well as the implications of regional coherence and reciprocity".

(3) [Kosovo has met the requirements of its visa liberalisation roadmap. On the basis of this assessment and taking account of all the criteria listed in Article -1 of Regulation (EC) No 539/2001, it is appropriate to exempt persons from Kosovo from the visa requirement when travelling to the territory of the Member States.]

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

(4) Kosovo should thus be transferred from Annex I, Part 2 to Regulation (EC) No 539/2001 to Annex II, Part 4 thereof. This visa waiver should apply only to holders of biometric passports issued in line with the standards of International Civil Aviation Organisation (ICAO) and Council Regulation (EC) No 2252/2004 [20].

(5) The visa exemption is dependent upon the continued implementation of the requirements of the visa liberalisation roadmap. The Commission will actively monitor the implementation of these requirements through the post-visa liberalisation mechanism. The visa exemption may be suspended by the EU in line with the suspension mechanism established by Article 1a of Regulation (EC) No 539/2001, as amended by Regulation xxx should the conditions set out therein be met.

(6) This Regulation constitutes a development of provisions of the Schengen acquis in which the United Kingdom does not take part, in accordance with Council Decision 2000/365/EC [21]. The United Kingdom is therefore not taking part in the adoption of this Regulation and is not bound by it or subject to its application.

(7) This Regulation constitutes a development of provisions of the Schengen acquis in which Ireland does not take part, in accordance with Council Decision 2002/192/EC [22] . Ireland is therefore not taking part in the adoption of this Regulation and is not bound by it or subject to its application.

(8) As regards Iceland and Norway, this Regulation constitutes a development of provisions of the Schengen acquis within the meaning of the Agreement concluded by the Council of the European Union and the Republic of Iceland and the Kingdom of Norway concerning the association of those two States with the implementation, application and development of the Schengen acquis, which fall within the area referred to in point B of Article 1, of Council Decision 1999/437/EC [23] .

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

(9) As regards Switzerland, this Regulation constitutes a development of the provisions of the Schengen acquis within the meaning of the Agreement between the European Union, the European Community and the Swiss Confederation on the Swiss Confederation's association with the implementation, application and development of the Schengen acquis, which fall within the area referred to in point B of Article 1, of Decision 1999/437/EC, read in conjunction with Article 3 of Council Decision 2008/146/EC [24] .

(10) As regards Liechtenstein, this Regulation constitutes a development of the provisions of the Schengen acquis within the meaning of the Protocol signed between the European Union, the European Community, the Swiss Confederation and the Principality of Liechtenstein on the accession of the Principality of Liechtenstein to the Agreement between the European Union, the European Community and the Swiss Confederation on the Swiss Confederation's association with the implementation, application and development of the Schengen acquis, which fall within the area referred to in point B of Article 1, of Decision 1999/437/EC read in conjunction with Article 3 of Council Decision 2011/350/EU [25] ,

HAVE ADOPTED THIS REGULATION:

Article 1

Regulation (EC) No 539/2001 is amended as follows:

a)in Annex I, Part 2 ("ENTITIES AND TERRITORIAL AUTHORITIES THAT ARE NOT RECOGNISED AS STATES BY AT LEAST ONE MEMBER STATE"), the reference to Kosovo as defined by the United Nations Security Council Resolution 1244 of 10 June 1999 is deleted.

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

(b)in Annex II, Part 4 ("ENTITIES AND TERRITORIAL AUTHORITIES THAT ARE NOT RECOGNISED AS STATES BY AT LEAST ONE MEMBER STATE"), the following reference is inserted:

Article 2

This Regulation shall enter into force on the twentieth day following that of its publication in the Official Journal of the European Union.

This Regulation shall be binding in its entirety and directly applicable in the Member States in accordance with the Treaties.

Done at Brussels,

Figure A.1. Visa Regulation of the European Parliament and of the Council (cont.).

A.2. MENOMINEE INDIAN TRIBE OF WISCONSIN v. UNITED STATES ET AL.

SUPREME COURT OF THE UNITED STATES

MENOMINEE INDIAN TRIBE OF WISCONSIN v. UNITED STATES ET AL.

CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE DISTRICT OF COLUMBIA CIRCUIT

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al..

No. 14–510. Argued December 1, 2015—Decided January 25, 2016

Pursuant to the Indian Self-Determination and Education Assistance Act (ISDA), petitioner Menominee Indian Tribe of Wisconsin contracted with the Indian Health Service (IHS) to operate what would otherwise have been a federal program and to receive an amount of money equal to what the Government would have spent on operating the program itself, including reimbursement for reasonable contract support costs. 25 U. S. C. §§450f, 450j–1(a). After other tribal entities successfully litigated complaints against the Federal Government for failing to honor its obligation to pay contract support costs, the Menominee Tribe presented its own contract support claims to the IHS in accordance with the Contract Disputes Act of 1978 (CDA), which requires contractors to present each claim to a contracting officer for decision, 41 U. S. C. §7103(a)(1). The contracting officer denied some of the Tribe's claims because they were not presented within the CDA's 6-year limitations period. See §7103(a)(4)(A). The Tribe challenged the denials in Federal District Court, arguing that the limitations period should be tolled for the nearly two years in which a putative class action, brought by tribes with parallel complaints, was pending. As relevant here, the District Court eventually denied the Tribe's equitable-tolling claim, and the Court of Appeals affirmed, holding that no extraordinary circumstances beyond the Tribe's control caused the delay.

Held: Equitable tolling does not apply to the presentment of petitioner's claims. Pp. 5–9.

(a) To be entitled to equitable tolling of a statute of limitations, a litigant must establish "(1) that he has been pursuing his rights diligently, and (2) that some extraordinary circumstance stood in his way and prevented timely filing." Holland v. Florida, 560 U. S. 631, 649. The Tribe argues that diligence and extraordinary circumstances should be considered together as factors in a unitary test, and it

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

faults the Court of Appeals for declining to consider the Tribe's diligence in connection with its finding that no extraordinary circumstances existed. But this Court has expressly characterized these two components as "elements" not merely factors of indeterminate or commensurable weight, Pace v. DiGuglielmo, 544 U. S. 408, 418, and has treated them as such in practice, see Lawrence v. Florida, 549 U. S. 327, 336–337. The Tribe also objects to the Court of Appeals' interpretation of the "extraordinary circumstances" prong as requiring the showing of an "external obstacle" to timely filing. This Court reaffirms that this prong is met only where the circumstances that caused a litigant's delay are both extraordinary and beyond its control. Pp. 5–7.

(b) None of the Tribe's excuses satisfy the "extraordinary circumstances" prong of the test. The Tribe had unilateral authority to present its claims in a timely manner. Its claimed obstacles, namely, a mistaken reliance on a putative class action and a belief that presentment was futile, were not outside the Tribe's control. And the significant risk and expense associated with presenting and litigating its claims are far from extraordinary. Finally, the special relationship between the United States and Indian tribes, as articulated in the ISDA, does not override clear statutory language. Pp. 7–8. 764 F. 3d 51, affirmed.

ALITO, J., delivered the opinion for a unanimous Court.

JUSTICE ALITO delivered the opinion of the Court. Petitioner Menominee Indian Tribe of Wisconsin (Tribe) seeks equitable tolling to preserve contract claims not timely presented to a federal contracting officer. Because the Tribe cannot establish extraordinary circumstances that stood in the way of timely filing, we hold that equitable tolling does not apply.

Congress enacted the Indian Self-Determination and Education Assistance Act,

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

Pub. L. 93–638, 88 Stat. 2203, 25 U. S. C. §450 et seq., in 1975 to help Indian tribes assume responsibility for aid programs that benefit their members. Under the ISDA, tribes may enter into "selfdetermination contracts" with federal agencies to take control of a variety of federally funded programs §450f. A contracting tribe is eligible to receive the amount of money that the Government would have otherwise spent on the program, see §450j–1(a)(1), as well as reimbursement for reasonable "contract support costs," which include administrative and overhead costs associated with carrying out the contracted programs, §§450j–1(a)(2), (3), (5).

In 1988, Congress amended the ISDA to apply the Contract Disputes Act of 1978 (CDA), 41 U. S. C. §7101 et seq., to disputes arising under the ISDA. See 25 U. S. C. §450m–1(d); Indian Self-Determination and Education Assistance Act Amendments of 1988, §206(2), 102 Stat. 2295. As part of its mandatory administrative process for resolving contract disputes, the CDA requires contractors to present "[e]ach claim" they may have to a contracting officer for decision. 41 U. S. C. §7103(a)(1). Congress later amended the CDA to include a 6-year statute of limitations for presentment of each claim. Federal Acquisition Streamlining Act of 1994, 41 U. S. C. §7103(a)(4)(A). Under the CDA, the contracting officer's decision is generally final, unless challenged through one of the statutorily authorized routes. §7103(g). A contractor dissatisfied with the officer's decision may either take an administrative appeal to a board of contract appeals or file an action for breach of contract in the United States Court of Federal Claims. §§7104(a), (b)(1), 7105(b). Both routes then lead to the United States Court of Appeals for the Federal Circuit for any further review. 28 U. S. C. §1295(a)(3); 41 U. S. C. §7107(a)(1); see 25 U. S. C. §450m–1(d). Under the ISDA, tribal contractors have a third option. They may file a claim for money damages in federal district court, §§450m–1(a), (d), and if they lose, they may pursue an appeal in one of the regional courts of appeals, 28 U. S. C. §1291.

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

Tribal contractors have repeatedly complained that the Federal Government has not fully honored its obligations to pay contract support costs. Three lawsuits making such claims are relevant here.

The first was a class action filed by the Ramah Navajo Chapter alleging that the Bureau of Indian Affairs (BIA) systematically underpaid certain contract support costs. Ramah Navajo Chapter v. Lujan, No. 1:90–cv–0957 (D NM) (filed Oct. 4, 1990). In 1993, Ramah successfully moved for certification of a nationwide class of all tribes that had contracted with the BIA under the ISDA. See Order and Memorandum Opinion in Ramah Navajo Chapter v. Lujan, No. 1:90–cv–0957 (D NM, Oct. 1, 1993), App. 35–40. The Government argued that each tribe needed to present its claims to a contracting officer before it could participate in the class. Id., at 37–38. But the trial court held that tribal contractors could participate in the class without presentment, because the suit alleged systemwide flaws in the BIA's contracting scheme, not merely breaches of individual contracts. Id., at 39. The Government did not appeal the certification order, and the Ramah class action proceeded to further litigation and settlement. The second relevant ISDA suit raised similar claims about contract support costs but arose from contracts with the Indian Health Service (IHS). Cherokee Nation of Okla. v. United States, No. 6:99–cv–0092 (ED Okla.) (filed Mar. 5, 1999). In Cherokee Nation, two tribes filed a putative class action against IHS. On February 9, 2001, the District Court denied class certification without addressing whether tribes would need to present claims to join the class. Cherokee Nation of Okla. v. United States, 199 F. R. D. 357, 363–366 (ED Okla.). The two plaintiff tribes did not appeal the denial of class certification but proceeded to the merits on their own, eventually prevailing before this Court in a parallel suit. See Cherokee Nation of Okla. v. Leavitt, 543 U. S. 631 (2005).

The third relevant case is the one now before us. In this case, the Tribe presented

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

its contract support claims (for contract years 1995 through 2004) to IHS on September 7, 2005, shortly after our Cherokee Nation ruling. As relevant here, the contracting officer denied the Tribe's claims based on its 1996, 1997, and 1998 contracts because, inter alia, those claims were barred by the CDA's 6-year statute of limitations.[1]

The Tribe challenged the denials in the United States District Court for the District of Columbia, arguing, based on theories of class-action and equitable tolling, that the limitations period should be tolled for the 707 days that the putative Cherokee Nation class had been pending. See American Pipe & Constr. Co. v. Utah, 414 U. S. 538 (1974) (class-action tolling); Holland v. Florida, 560 U. S. 631 (2010) (equitable tolling).

Initially, the District Court held that the limitations period was jurisdictional and thus forbade tolling of any sort. 539 F. Supp. 2d 152, 154, and n. 2 (DDC 2008). On appeal, the United States Court of Appeals for the District of Columbia Circuit concluded that the limitations period was not jurisdictional and thus did not necessarily bar tolling. 614 F. 3d 519, 526 (2010). But the court held that the Tribe was ineligible for class -action tolling during the pendency of the putative Cherokee Nation class, because the Tribe's failure to present its claims to IHS made it "ineligible to participate in the class action at the time class certification [was] denied." 614 F. 3d, at 527 (applying American Pipe). The court then remanded the case to the District Court to determine the Tribe's eligibility for equitable tolling.

On remand, the District Court concluded that the Tribe's asserted reasons for failing to present its claims within the specified time "do not, individually or collectively, amount to an extraordinary circumstance" that could warrant equitable tolling. 841 F. Supp. 2d 99, 107 (DC 2012) (internal quotation marks omitted).

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

This time, the Court of Appeals affirmed. 764 F. 3d 51 (CADC 2014). It explained that, "[t]o count as sufficiently 'extraordinary' to support equitable tolling, the circumstances that caused a litigant's delay must have been beyond its control," and "cannot be a product of that litigant's own misunderstanding of the law or tactical mistakes in litigation." Id., at 58. Because none of the Tribe's proffered circumstances was beyond its control, the court held, there were no extraordinary circumstances that could merit equitable tolling.

The Court of Appeals' decision created a split with the Federal Circuit, which granted another tribal entity equitable tolling under similar circumstances. See Arctic Slope Native Assn., Ltd. v. Sebelius, 699 F. 3d 1289 (CA Fed. 2012). We granted certiorari to resolve the conflict. 576. U. S.(2015).

The Court of Appeals denied the Tribe's request for equitable tolling by applying the test that we articulated in Holland v. Florida, 560 U. S. 631. Under Holland, a litigant is entitled to equitable tolling of a statute of limitations only if the litigant establishes two elements: "(1) that he has been pursuing his rights diligently, and (2) that some extraordinary circumstance stood in his way and prevented timely filing." Id., at 649 (internal quotation marks omitted).

The Tribe calls this formulation of the equitable tolling test overly rigid, given the doctrine's equitable nature. First, it argues that diligence and extraordinary circumstances should be considered together as two factors in a unitary test, and it faults the Court of Appeals for declining to consider the Tribe's diligence in connection with its finding that no extraordinary circumstances existed. But we have expressly characterized equitable tolling's two components as "elements," not merely factors of indeterminate or commensurable weight. Pace v. DiGuglielmo, 544 U. S. 408, 418 (2005) ("Generally, a litigant seeking equitable tolling bears the burden of establishing two elements"). And we have treated the two requirements

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

as distinct elements in practice, too, rejecting requests for equitable tolling where a litigant failed to satisfy one without addressing whether he satisfied the other. See, e.g., Lawrence v. Florida, 549 U. S. 327, 336–337 (2007) (rejecting equitable tolling without addressing diligence because habeas petitioner fell "far short of showing 'extraordinary circumstances'") Pace, supra, at 418 (holding, without resolving litigant's argument that he had "satisfied the extraordinary circumstance test," that, "[e]ven if we were to accept [his argument], he would not be entitled to relief because he has not established the requisite diligence").

Second, the Tribe objects to the Court of Appeals' interpretation of the "extraordinary circumstances" prong as requiring a litigant seeking tolling to show an "external obstacl[e]" to timely filing, i.e., that "the circumstances that caused a litigant's delay must have been beyond its control." 764 F. 3d, at 58–59. The Tribe complains that this "external obstacle" formulation amounts to the same kind of "'overly rigid per se approach'" we rejected in Holland. Brief for Petitioner 32 (quoting 560 U. S., at 653). But in truth, the phrase "external obstacle" merely reflects our requirement that a litigant seeking tolling show "that some extraordinary circumstance stood in his way." Id., at 649 (emphasis added; internal quotation marks omitted). This phrasing in Holland (and in Pace before that) would make little sense if equitable tolling were available when a litigant was responsible for its own delay. Indeed, the diligence prong already covers those affairs within the litigant's control; the extraordinarycircumstances prong, by contrast, is meant to cover matters outside its control. We therefore reaffirm that the second prong of the equitable tolling test is met only where the circumstances that caused a litigant's delay are both extraordinary and beyond its control.[2]

The Tribe offers no circumstances that meet this standard.

Its mistaken reliance on the putative Cherokee Nation class action was not an

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

obstacle beyond its control.[3] As the Tribe conceded below, see 614 F. 3d, at 526–527, it could not have been a member of the putative Cherokee Nation class because it did not present its claims to an IHS contracting officer before class certification was denied. Before then, the Tribe had unilateral authority to present its claims and to join the putative class. Presentment was blocked not by an obstacle outside its control, but by the Tribe's mistaken belief that presentment was unneeded.

The Tribe's mistake, in essence, was its inference that the reasoning of the Ramah class certification decision (allowing tribes to participate—without presentment—in the class challenging underpayment of BIA contract support costs) applied to the putative Cherokee Nation class. This mistake was fundamentally no different from "a garden variety claim of excusable neglect," Irwin v. Department of Veterans Affairs, 498 U. S. 89, 96 (1990), "such as a simple 'miscalculation' that leads a lawyer to miss a filing deadline," Holland, supra, at 651 (quoting Lawrence, supra, at 336). And it is quite different from relying on actually binding precedent that is subsequently reversed.[4]

The Tribe's other excuses are even less compelling. Its belief that presentment was futile was not an obstacle beyond its control but a species of the same mistake that kept it out of the putative Cherokee Nation class. And the fact that there may have been significant risk and expense associated with presenting and litigating its claims is far from extraordinary. As the District Court noted below, "it is common for a litigant to be confronted with significant costs to litigation, limited financial resources, an uncertain outcome based upon an uncertain legal landscape, and impending deadlines. These circumstances are not 'extraordinary.'" 841 F. Supp. 2d, at 107.

Finally, the Tribe also urges us to consider the special relationship between the

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

United States and Indian tribes, as articulated in the ISDA. See 25 U. S. C. §450a(b) ("Congress declares its commitment to the maintenance of the Federal Government's unique and continuing relationship with, and responsibility to, individual Indian tribes and to the Indian people as a whole"). We do not question the "general trust relationship between the United States and the Indian tribes," but any specific obligations the Government may have under that relationship are "governed by statute rather than the common law." United States v. Jicarilla Apache Nation, 564 U. S. 162, 165 (2011). The ISDA and CDA establish a clear procedure for the resolution of disputes over ISDA contracts, with an unambiguous 6-year deadline for presentment of claims. The "general trust relationship" does not override the clear language of those statutes.[5] For these reasons, the judgment of the United States Court of Appeals for the District of Columbia Circuit is affirmed.

It is so ordered.

Figure A.2. Menominee Indian Tribe of Wisconsin v. United States et al. (cont.).

A.3. THE NICE PEOPLE

"They certainly are nice people,"assented to my wife's observation, using the colloquial phrase with a consciousness that it was anything but "nice" English, "and I'll bet that their three children are better brought up than most of—-"

"Two children," corrected my wife.

"Three, he told me."

"My dear, she said there were two."

"He said three."

"You've simply forgotten. I'm sure she told me they had only two–a boy and a girl."

"Well, I didn't enter into particulars."

"No, dear, and you couldn't have understood him. Two children."

"All right," said; but I did not think it was all right. As a near-sighted man learns by enforced observation to recognize persons at a distance when the face is not

Figure A.3. The Nice People by Henry Cuyler Bunner.

visible to the normal eye, so the man with a bad memory learns, almost unconsciously, to listen carefully and report accurately. My memory is bad; but I had not had time to forget that Mr. Brewster Brede had told me that afternoon that he had three children, at present left in the care of his mother-in-law, while he and Mrs. Brede took their summer vacation.

"Two children," repeated my wife; "and they are staying with his aunt Jenny."

"He told me with his mother-in-law," put in. My wife looked at me with a serious expression. Men may not remember much of what they are told about children; but any man knows the difference between an aunt and a mother-in-law.

"But don't you think they're nice people?" asked my wife.

"Oh, certainly," replied."Only they seem to be a little mixed up about their children."

"That isn't a nice thing to say," returned my wife. I could not deny it.

And yet, the next morning, when the Bredes came down and seated themselves opposite us at table, beaming and smiling in their natural, pleasant, well-bred fashion, I knew, to a social certainty, that they were "nice" people. He was a fine-looking fellow in his neat tennis-flannels, slim, graceful, twenty-eight or thirty years old, with a Frenchy pointed beard. She was "nice" in all her pretty clothes, and she herself was pretty with that type of prettiness which outwears most other types–the prettiness that lies in a rounded figure, a dusky skin, plump, rosy cheeks, white teeth and black eyes. She might have been twenty-five; you guessed that she was prettier than she was at twenty, and that she would be prettier still at forty.

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

And nice people were all we wanted to make us happy in Mr. Jacobus's summer boarding-house on top of Orange Mountain. For a week we had come down to breakfast each morning, wondering why we wasted the precious days of idleness with the company gathered around the Jacobus board. What joy of human companionship was to be had out of Mrs. Tabb and Miss Hoogencamp, the two middle-aged gossips from Scranton, Pa.–out of Mr. and Mrs. Biggle, an indurated head-bookkeeper and his prim and censorious wife–out of old Major Halkit, a retired business man, who, having once sold a few shares on commission, wrote for circulars of every stock company that was started, and tried to induce every one to invest who would listen to him? We looked around at those dull faces, the truthful indices of mean and barren minds, and decided that we would leave that morning. Then we ate Mrs. Jacobus's biscuit, light as Aurora's cloudlets, drank her honest coffee, inhaled the perfume of the late azaleas with which she decked her table, and decided to postpone our departure one more day. And then we wandered out to take our morning glance at what we called "our view" and it seemed to us as if Tabb and Hoogencamp and Halkit and the Biggleses could not drive us away in a year.

I was not surprised when, after breakfast, my wife invited the Bredes to walk with us to "our view." The Hoogencamp-Biggle-Tabb-Halkit contingent never stirred off Jacobus's veranda; but we both felt that the Bredes would not profane that sacred scene. We strolled slowly across the fields, passed through the little belt of woods and, as I heard Mrs. Brede's little cry of startled rapture, I motioned to Brede to look up.

"By Jove!" he cried,"heavenly!"

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

We looked off from the brow of the mountain over fifteen miles of billowing green, to where, far across a far stretch of pale blue lay a dim purple line that we knew was Staten Island. Towns and villages lay before us and under us; there were ridges and hills, uplands and lowlands, woods and plains, all massed and mingled in that great silent sea of sunlit green. For silent it was to us, standing in the silence of a high place–silent with a Sunday stillness that made us listen, without taking thought, for the sound of bells coming up from the spires that rose above the tree-tops–the tree-tops that lay as far beneath us as the light clouds were above us that dropped great shadows upon our heads and faint specks of shade upon the broad sweep of land at the mountain's foot.

"And so that is your view?" asked Mrs. Brede, after a moment; "you are very generous to make it ours, too."

Then we lay down on the grass, and Brede began to talk, in a gentle voice, as if he felt the influence of the place. He had paddled a canoe, in his earlier days, he said, and he knew every river and creek in that vast stretch of landscape. He found his landmarks, and pointed out to us where the Passaic and the Hackensack flowed, invisible to us, hidden behind great ridges that in our sight were but combings of the green waves upon which we looked down. And yet, on the further side of those broad ridges and rises were scores of villages–a little world of country life, lying unseen under our eyes.

"A good deal like looking at humanity," he said; "there is such a thing as getting so far above our fellow men that we see only one side of them."

Ah, how much better was this sort of talk than the chatter and gossip of the Tabb and the Hoogencamp–than the Major's dissertations upon his everlasting circulars! My wife and I exchanged glances.

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

"Now, when I went up the Matterhorn" Mr. Brede began.

"Why, dear," interrupted his wife, "I didn't know you ever went up the Matterhorn."

"It–it was five years ago,"said Mr. Brede, hurriedly. "I–I didn't tell you–when I was on the other side, you know–it was rather dangerous–well, as I was saying–it looked–oh, it didn't look at all like this."

A cloud floated overhead, throwing its great shadow over the field where we lay. The shadow passed over the mountain's brow and reappeared far below, a rapidly decreasing blot, flying eastward over the golden green. My wife and I exchanged glances once more.

Somehow, the shadow lingered over us all. As we went home, the Bredes went side by side along the narrow path, and my wife and I walked together.

"Should you think," she asked me, "that a man would climb the Matterhorn the very first year he was married?"

"I don't know, my dear," answered, evasively; "this isn't the first year I have been married, not by a good many, and I wouldn't climb it–for a farm."

"You know what I mean," she said.

I did.

When we reached the boarding-house, Mr. Jacobus took me aside.

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

"You know," he began his discourse, "my wife she uset to live in N' York!"

I didn't know, but I said "Yes."

"She says the numbers on the streets runs criss-cross-like. Thirty-four's on one side o' the street an' thirty-five on t'other. How's that?"

"That is the invariable rule, I believe."

"Then–I say–these here new folk that you 'n' your wife seem so mighty taken up with–d'ye know anything about 'em?"

"I know nothing about the character of your boarders, Mr. Jacobus," replied, conscious of some irritability. "If I choose to associate with any of them—-"

"Jess so–jess so!" broke in Jacobus. "I hain't nothin' to say ag'inst yer sosherbil'ty. But do ye know them?"

"Why, certainly not," replied.

"Well–that was all I wuz askin' ye. Ye see, when he come here to take the rooms–you wasn't here then–he told my wife that he lived at number thirty-four in his street. An' yistiddy she told her that they lived at number thirty-five. He said he lived in an apartment-house. Now there can't be no apartment-house on two sides of the same street, kin they?"

"What street was it?" inquired, wearily.

"Hundred 'n' twenty-first street."

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

"May be," replied, still more wearily. "That's Harlem. Nobody knows what people will do in Harlem."

I went up to my wife's room.

"Don't you think it's queer?" she asked me.

"I think I'll have a talk with that young man to-night," said, "and see if he can give some account of himself."

"But, my dear," my wife said, gravely, "she doesn't know whether they've had the measles or not."

"Why, Great Scott!" exclaimed,"they must have had them when they were children."

"Please don't be stupid," said my wife. "I meant their children."

After dinner that night–or rather, after supper, for we had dinner in the middle of the day at Jacobus's–I walked down the long verandah to ask Brede, who was placidly smoking at the other end, to accompany me on a twilight stroll. Half way down I met Major Halkit.

"That friend of yours," he said, indicating the unconscious figure at the further end of the house, "seems to be a queer sort of a Dick. He told me that he was out of business, and just looking round for a chance to invest his capital. And I've been telling him what an everlasting big show he had to take stock in the

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

Capitoline Trust Company–starts next month–four million capital–I told you all about it. 'Oh, well,' he says, 'let's wait and think about it.' 'Wait!' says I, 'the Capitoline Trust Company won't wait for you, my boy. This is letting you in on the ground floor,' says I, 'and it's now or never.' 'Oh, let it wait,' says he. I don't know what's in-to the man."

"I don't know how well he knows his own business, Major," said as I started again for Brede's end of the veranda. But I was troubled none the less. The Major could not have influenced the sale of one share of stock in the Capitoline Company. But that stock was a great investment; a rare chance for a purchaser with a few thousand dollars. Perhaps it was no more remarkable that Brede should not invest than that I should not–and yet, it seemed to add one circumstance more to the other suspicious circumstances. When I went upstairs that evening, I found my wife putting her hair to bed–I don't know how I can better describe an operation familiar to every married man. I waited until the last tress was coiled up, and then I spoke: "I've talked with Brede," said,"and I didn't have to catechize him. He seemed to feel that some sort of explanation was looked for, and he was very outspoken. You were right about the children–that is, I must have misunderstood him. There are only two. But the Matterhorn episode was simple enough. He didn't realize how dangerous it was until he had got so far into it that he couldn't back out; and he didn't tell her, because he'd left her here, you see, and under the circumstances—-"

"Left her here!" cried my wife."I've been sitting with her the whole afternoon, sewing, and she told me that he left her at Geneva, and came back and took her to Basle, and the baby was born there–now I'm sure, dear, because I asked her."

"Perhaps I was mistaken when I thought he said she was on this side of the water," suggested, with bitter, biting irony.

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

"You poor dear, did I abuse you?" said my wife. "But, do you know, Mrs. Tabb said that she didn't know how many lumps of sugar he took in his coffee. Now that seems queer, doesn't it?"

It did. It was a small thing. But it looked queer, Very queer. The next morning, it was clear that war was declared against the Bredes. They came down to breakfast somewhat late, and, as soon as they arrived, the Biggleses swooped up the last fragments that remained on their plates, and made a stately march out of the dining-room, Then Miss Hoogencamp arose and departed, leaving a whole fish-ball on her plate. Even as Atalanta might have dropped an apple behind her to tempt her pursuer to check his speed, so Miss Hoogencamp left that fish-ball behind her, and between her maiden self and contamination.

We had finished our breakfast, my wife and I, before the Bredes appeared. We talked it over, and agreed that we were glad that we had not been obliged to take sides upon such insufficient testimony.

After breakfast, it was the custom of the male half of the Jacobus household to go around the corner of the building and smoke their pipes and cigars where they would not annoy the ladies. We sat under a trellis covered with a grapevine that had borne no grapes in the memory of man. This vine, however, bore leaves, and these, on that pleasant summer morning, shielded from us two persons who were in earnest conversation in the straggling, half-dead flower-garden at the side of the house.

"I don't want," we heard Mr. Jacobus say, "to enter in no man's pry-vacy; but I do want to know who it may be, like, that I hev in my house. Now what I ask of you, and I don't want you to take it as in no ways personal, is–hev you your merridge-license with you?"

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

I think it was a chance shot; but it told all the same. The Major (he was a widower) and Mr. Biggle and I looked at each other; and Mr. Jacobus, on the other side of the grape-trellis, looked at–I don't know what–and was as silent as we were. Where is your marriage-license, married reader? Do you know? Four men, not including Mr. Brede, stood or sat on one side or the other of that grape-trellis, and not one of them knew where his marriage-license was. Each of us had had one–the Major had had three. But where were they? Where is yours? Tucked in your best-man's pocket; deposited in his desk–or washed to a pulp in his white waistcoat (if white waistcoats be the fashion of the hour), washed out of existence–can you tell where it is? Can you–unless you are one of those people who frame that interesting document and hang it upon their drawing-room walls?

Mr. Brede's voice arose, after an awful stillness of what seemed like five minutes, and was, probably, thirty seconds:

"Mr. Jacobus, will you make out your bill at once, and let me pay it? I shall leave by the six o'clock train. And will you also send the wagon for my trunks?"

"I hain't said I wanted to hev ye leave—-" began Mr. Jacobus; but Brede cut him short.

"Bring me your bill."

"But," remonstrated Jacobus,"ef ye ain't—-"

"Bring me your bill!" said Mr. Brede.

My wife and I went out for our morning's walk. But it seemed to us, when we looked at "our view," as if we could only see those invisible villages of which Brede

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

had told us–that other side of the ridges and rises of which we catch no glimpse from lofty hills or from the heights of human self-esteem. We meant to stay out until the Bredes had taken their departure; but we returned just in time to see Pete, the Jacobus darkey, the blacker of boots, the brasher of coats, the general handy-man of the house, loading the Brede trunks on the Jacobus wagon.

And, as we stepped upon the verandah, down came Mrs. Brede, leaning on Mr. Brede's arm, as though she were ill; and it was clear that she had been crying. There were heavy rings about her pretty black eyes.

My wife took a step toward her.

"Look at that dress, dear," she whispered; "she never thought anything like this was going to happen when she put that on."

It was a pretty, delicate, dainty dress, a graceful, narrow-striped affair. Her hat was trimmed with a narrow-striped silk of the same colors–maroon and white–and in her hand she held a parasol that matched her dress.

"She's had a new dress on twice a day," said my wife, "but that's the prettiest yet. Oh, somehow–I'm awfully sorry they're going!"

But going they were. They moved toward the steps. Mrs. Brede looked toward my wife, and my wife moved toward Mrs. Brede. But the ostracized woman, as though she felt the deep humiliation of her position, turned sharply away, and opened her parasol to shield her eyes from the sun. A shower of rice–a half-pound shower of rice–fell down over her pretty hat and her pretty dress, and fell in a spattering circle on the floor, outlining her skirts–and there it lay in a broad, uneven band, bright in the morning sun.

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

Mrs. Brede was in my wife's arms, sobbing as if her young heart would break.

"Oh, you poor, dear, silly children!" my wife cried, as Mrs. Brede sobbed on her shoulder, "why didn't you tell us?"

"W-W-W-We didn't want to be t-t-taken for a b-b-b-b-bridal couple," sobbed Mrs. Brede;"and we d-d-didn't dream what awful lies we'd have to tell, and all the aw-awful mixed-up-ness of it. Oh, dear, dear, dear!"

"Pete!" commanded Mr. Jacobus, "put back them trunks. These folks stays here's long's they wants ter. Mr. Brede "–he held out a large, hard hand–" I'd orter've known better," he said. And my last doubt of Mr. Brede vanished as he shook that grimy hand in manly fashion.

The two women were walking off toward "our view," each with an arm about the other's waist–touched by a sudden sisterhood of sympathy.

"Gentlemen," said Mr. Brede, addressing Jacobus, Biggle, the Major and me, "there is a hostelry down the street where they sell honest New Jersey beer. I recognize the obligations of the situation."

We five men filed down the street. The two women went toward the pleasant slope where the sunlight gilded the forehead of the great hill. On Mr. Jacobus's veranda lay a spattered circle of shining grains of rice. Two of Mr. Jacobus's pigeons flew down and picked up the shining grains, making grateful noises far down in their throats.

Figure A.3. The Nice People by Henry Cuyler Bunner (cont.).

A.4. POLITICS AND THE ENGLISH LANGUAGE

Most people who bother with the matter at all would admit that the English language is in a bad way, but it is generally assumed that we cannot by conscious action do anything about it. Our civilization is decadent, and our language–so the argument runs–must inevitably share in the general collapse. It follows that any struggle against the abuse of language is a sentimental archaism, like preferring candles to electric light or hansom cabs to aeroplanes. Underneath this lies the half-conscious belief that language is a natural growth and not an instrument which we shape for our own purposes.

Now, it is clear that the decline of a language must ultimately have political and economic causes: it is not due simply to the bad influence of this or that individual writer. But an effect can become a cause, reinforcing the original cause and producing the same effect in an intensified form, and so on indefinitely. A man may take to drink because he feels himself to be a failure, and then fail all the more completely because he drinks. It is rather the same thing that is happening to the English language. It becomes ugly and inaccurate because our thoughts are foolish, but the slovenliness of our language makes it easier for us to have foolish thoughts. The point is that the process is reversible. Modern English, especially written English, is full of bad habits which spread by imitation and which can be avoided if one is willing to take the necessary trouble. If one gets rid of these habits one can think more clearly, and to think clearly is a necessary first step towards

Figure A.4. Politics and the English language by George Orwell.

political regeneration: so that the fight against bad English is not frivolous and is not the exclusive concern of professional writers. I will come back to this presently, and I hope that by that time the meaning of what I have said here will have become clearer. Meanwhile, here are five specimens of the English language as it is now habitually written.

These five passages have not been picked out because they are especially bad–I could have quoted far worse if I had chosen–but because they illustrate various of the mental vices from which we now suffer. They are a little below the average, but are fairly representative samples. I number them so that I can refer back to them when necessary:

(1) I am not, indeed, sure whether it is not true to say that the Milton who once seemed not unlike a seventeenth-century Shelley had not become, out of an experience ever more bitter in each year, more alien (sic) to the founder of that Jesuit sect which nothing could induce him to tolerate.

PROFESSOR HAROLD LASKI (Essay in Freedom of Expression)

(2) Above all, we cannot play ducks and drakes with a native battery of idioms which prescribes such egregious collocations of vocables as the Basic put up with for tolerate or put at a loss for bewilder.

PROFESSOR LANCELOT HOGBEN (Interglossa)

(3) On the one side we have the free personality; by definition it is not neurotic, for it has neither conflict nor dream. Its desires, such as they are, are transparent, for they are just what institutional approval keeps in the forefront of consciousness; another institutional pattern would alter their number and intensity; there is little

Figure A.4. Politics and the English language by George Orwell (cont.).

in them that is natural, irreducible, or culturally dangerous. But on the other side, the social bond itself is nothing but the mutual reflection of these self-secure integrities. Recall the definition of love. Is not this the very picture of a small academic? Where is there a place in this hall of mirrors for either personality or fraternity?

ESSAY ON PSYCHOLOGY in Politics (New York)

(4) All the "best people" from the gentlemen's clubs, and all the frantic fascist captains, united in common hatred of Socialism and bestial horror of the rising tide of the mass revolutionary movement, have turned to acts of provocation, to foul incendiarism, to medieval legends of poisoned wells, to legalize their own destruction of proletarian organizations, and rouse the agitated petty-bourgeoisie to chauvinistic fervor on behalf of the fight against the revolutionary way out of the crisis.

COMMUNIST PAMPHLET

(5) If a new spirit is to be infused into this old country, there is one thorny and contentious reform which must be tackled, and that is the humanization and galvanization of the B.B.C. Timidity here will bespeak canker and atrophy of the soul. The heart of Britain may lee sound and of strong beat, for instance, but the British lion's roar at present is like that of Bottom in Shakespeare's Midsummer Night's Dream–as gentle as any sucking dove. A virile new Britain cannot continue indefinitely to be traduced in the eyes, or rather ears, of the world by the effete languors of Langham Place, brazenly masquerading as "standard English." When the Voice of Britain is heard at nine o'clock, better far and infinitely less ludicrous to hear aitches honestly dropped than the present priggish, inflated, inhibited, school-ma'am-ish arch braying of blameless bashful mewing maidens.

Figure A.4. Politics and the English language by George Orwell (cont.).

LETTER IN Tribune

Each of these passages has faults of its own, but quite apart from avoidable ugliness, two qualities are common to all of them. The first is staleness of imagery; the other is

lack of precision. The writer either has a meaning and cannot express it, or he inadvertently says something else, or he is almost indifferent as to whether his words mean anything or not. This mixture of vagueness and sheer incompetence is the most marked characteristic of modern English prose, and especially of any kind of political writing. As soon as certain topics are raised, the concrete melts into the abstract and no one seems able to think of turns of speech that are not hackneyed: prose consists less and less of words chosen for the sake of their meaning, and more and more of phrases tacked together like the sections of a prefabricated hen-house. I list below, with notes and examples, various of the tricks by means of which the work of prose-construction is habitually dodged:

Dying metaphors. A newly-invented metaphor assists thought by evoking a visual image, while on the other hand a metaphor which is technically "dead" (e.g., iron resolution) has in effect reverted to being an ordinary word and can generally be used without loss of vividness. But in between these two classes there is a huge dump of worn-out metaphors which have lost all evocative power and are merely used because they save people the trouble of inventing phrases for themselves. Examples are: Ring the changes on, take up the cudgels for, toe the line, ride roughshod over, stand shoulder to shoulder with, play into the hands of, an axe to grind, grist to the mill, fishing in troubled waters, on the order of the day, Achilles' heel, swan song, hotbed. Many of these are used without knowledge of their meaning (what is a "rift," for instance?), and incompatible metaphors are frequently mixed, a sure sign that the writer is not interested in what he is saying. Some metaphors now current have been twisted out of their original meaning

Figure A.4. Politics and the English language by George Orwell (cont.).

without those who use them even being aware of the fact. For example, toe the line is sometimes written tow the line. Another example is the hammer and the anvil, now always used with the implication that the anvil gets the worst of it. In real life it is always the anvil that breaks the hammer, never the other way about: a writer who stopped to think what he was saying would be aware of this, and would avoid perverting the original phrase.

Operators, or verbal false limbs. These save the trouble of picking out appropriate verbs and nouns, and at the same time pad each sentence with extra syllables which give it an appearance of symmetry. Characteristic phrases are: render inoperative, militate against, prove unacceptable, make contact with, be subjected to, give rise to, give grounds for, having the effect of, play a leading part (role) in, make itself felt, take effect, exhibit a tendency to, serve the purpose of, etc., etc. The keynote is the elimination of simple verbs. Instead of being a single word, such as break, stop, spoil, mend, kill, a verb becomes a phrase, made up of a noun or adjective tacked on to some general-purposes verb as prove, serve, form, play, render. In addition, the passive voice is wherever possible used in preference to the active, and noun constructions are used instead of gerunds (by examination of instead of by examining). The range of verbs is further cut down by means of the -ize and de- formations, and banal statements are given an appearance of profundity by means of the not un- formation. Simple conjunctions and prepositions are replaced by such phrases as with respect to, having regard to, the fact that, by dint of, in view of, in the interests of, on the hypothesis that; and the ends of sentences are saved from anti-climax by such resounding commonplaces as greatly to be desired, cannot be left out of account, a development to be expected in the near future, deserving of serious consideration, brought to a satisfactory conclusion, and so on and so forth.

Pretentious diction. Words like phenomenon, element, individual (as noun), ob-

Figure A.4. Politics and the English language by George Orwell (cont.).

jective, categorical, effective, virtual, basis, primary, promote, constitute, exhibit, exploit, utilize, eliminate, liquidate, are used to dress up simple statements and give an air of scientific impartiality to biased judgments. Adjectives like epoch-making, epic, historic, unforgettable, triumphant, age-old, inevitable, inexorable, veritable, are used to dignify the sordid processes of international politics, while writing that aims at glorifying war usually takes on an archaic color, its characteristic words being: realm, throne, chariot, mailed fist, trident, sword, shield, buckler, banner, jackboot, clarion. Foreign words and expressions such as cul de sac, ancien regime, deus ex machina, mutatis mutandis, status quo, gleichschaltung, weltanschauung, are used to give an air of culture and elegance. Except for the useful abbreviations i.e., e.g., and etc., there is no real need for any of the hundreds of foreign phrases now current in English. Bad writers, and especially scientific, political and sociological writers, are nearly always haunted by the notion that Latin or Greek words are grander than Saxon ones, and unnecessary words like expedite, ameliorate, predict, extraneous, deracinated, clandestine, subaqueous and hundreds of others constantly gain ground from their Anglo-Saxon opposite numbers.1 The jargon peculiar to Marxist writing (hyena, hangman, cannibal, petty bourgeois, these gentry, lackey, flunkey, mad dog, White Guard, etc.) consists largely of words and phrases translated from Russian, German or French; but the normal way of coining a new word is to use a Latin or Greek root with the appropriate affix and, where necessary, the -ize formation. It is often easier to make up words of this kind (de-regionalize, impermissible, extramarital, non-fragmentary and so forth) than to think up the English words that will cover one's meaning. The result, in general, is an increase in slovenliness and vagueness.

1. An interesting illustration of this is the way in which the English flower names which were in use till very recently are being ousted by Greek ones, snap-dragon

Figure A.4. Politics and the English language by George Orwell (cont.).

becoming antirrhinum, forget-me-not becoming myosotis, etc. It is hard to see any practical reason for this change of fashion: it is probably due to an instinctive turning-away from the more homely word and a vague feeling that the Greek word is scientific.

Meaningless words. In certain kinds of writing, particularly in art criticism and literary criticism, it is normal to come across long passages which are almost completely lacking in meaning.2 Words like romantic, plastic, values, human, dead, sentimental, natural, vitality, as used in art criticism, are strictly meaningless, in the sense that they not only do not point to any discoverable object, but are hardly even expected to do so by the reader. When one critic writes, "The outstanding feature of Mr. X's work is its living quality," while another writes, "The immediately striking thing about Mr. X's work is its peculiar deadness" the reader accepts this as a simple difference of opinion. If words like black and white were involved, instead of the jargon words dead and living, he would see at once that language was being used in an improper way. Many political words are similarly abused. The word Fascism has now no meaning except in so far as it signifies "something not desirable." The words democracy, socialism, freedom, patriotic, realistic, justice, have each of them several different meanings which cannot be reconciled with one another. In the case of a word like democracy, not only is there no agreed definition, but the attempt to make one is resisted from all sides. It is almost universally felt that when we call a country democratic we are praising it: consequently the defenders of every kind of régime claim that it is a democracy, and fear that they might have to stop using the word if it were tied down to any one meaning. Words of this kind are often used in a consciously dishonest way. That is, the person who uses them has his own private definition, but allows his hearer to think he means something quite different. Statements like Marshal Pétain was a true patriot, The Soviet Press is the freest in the world, The Catholic Church is opposed to persecution, are almost always made with intent to deceive. Other

Figure A.4. Politics and the English language by George Orwell (cont.).

words used in variable meanings, in most cases more or less dishonestly, are: class, totalitarian, science, progressive, reactionary bourgeois, equality.

2. Example: "Comfort's catholicity of perception and image, strangely Whitmanesque in range, almost the exact opposite in aesthetic compulsion, continues to evoke that trembling atmospheric accumulative hinting at a cruel, an inexorably serene timelessness . . . Wrey Gardiner scores by aiming at simple bullseyes with precision. Only they are not so simple, and through this contented sadness runs more than the surface bittersweet of resignation." (Poetry Quarterly.)

Now that I have made this catalogue of swindles and perversions, let me give another example of the kind of writing that they lead to. This time it must of its nature be an imaginary one. I am going to translate a passage of good English into modern English of the worst sort. Here is a well-known verse from Ecclesiastes:

I returned, and saw under the sun, that the race is not to the swift, nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding, nor yet favor to men of skill; but time and chance happeneth to them all.

Here it is in modern English:

Objective consideration of contemporary phenomena compels the conclusion that success or failure in competitive activities exhibits no tendency to be commensurate with innate capacity, but that a considerable element of the unpredictable must invariably be taken into account.

This is a parody, but not a very gross one. Exhibit (3), above, for instance, contains several patches of the same kind of English. It will be seen that I have not made a full translation. The beginning and ending of the

Figure A.4. Politics and the English language by George Orwell (cont.).

sentence follow the original meaning fairly closely, but in the middle the concrete illustrations–race, battle, bread–dissolve into the vague phrase "success or failure in competitive activities"Ṫhis had to be so, because no modern writer of the kind I am discussing–no one capable of using phrases like objective consideration of contemporary phenomena" s–would ever tabulate his thoughts in that precise and detailed way. The whole tendency of modern prose is away from concreteness. Now analyze these two sentences a little more closely. The first contains 49 words but only 60 syllables, and all its words are those of everyday life. The second contains 38 words of 90 syllables: 18 of its words are from Latin roots, and one from Greek. The first sentence contains six vivid images, and only one phrase ("time and chance") that could be called vague. The second contains not a single fresh, arresting phrase, and in spite of its 90 syllables it gives only a shortened version of the meaning contained in the first. Yet without a doubt it is the second kind of sentence that is gaining ground in modern English. I do not want to exaggerate. This kind of writing is not yet universal, and outcrops of simplicity will occur here and there in the worst-written page. Still, if you or I were told to write a few lines on the uncertainty of human fortunes, we should probably come much nearer to my imaginary sentence than to the one from Ecclesiastes.

As I have tried to show, modern writing at its worst does not consist in picking out words for the sake of their meaning and inventing images in order to make the meaning clearer. It consists in gumming together long strips of words which have already been set in order by someone else, and making the results presentable by sheer humbug. The attraction of this way of writing, is that it is easy. It is easier–even quicker, once you have the habit–to say In my opinion it is a not unjustifiable assumption that than to say I think. If you use ready-made phrases, you not only don't have to hunt about for words; you also don't have to bother with the rhythms of your sentences, since these phrases are generally so arranged as to be more or less euphonious. When you are composing in a hurry–when you are dictating to

Figure A.4. Politics and the English language by George Orwell (cont.).

a stenographer, for instance, or making a public speech–it is natural to fall into a pretentious, Latinized style. Tags like a consideration which we should do well to bear in mind or a conclusion to which all of us would readily assent will save many a sentence from coming down with a bump. By using stale metaphors, similes and idioms, you save much mental effort at the cost of leaving your meaning vague, not only for your reader but for yourself. This is the significance of mixed metaphors. The sole aim of a metaphor is to call up a visual image. When these images clash– as in The Fascist octopus has sung its swan song, the jackboot is thrown into the melting pot–it can be taken as certain that the writer is not seeing a mental image of the objects he is naming; in other words he is not really thinking. Look again at the examples I gave at the beginning of this essay. Professor Laski (1) uses five negatives in 53 words. One of these is superfluous, making nonsense of the whole passage, and in addition there is the slip alien for akin, making further nonsense, and several avoidable pieces of clumsiness which increase the general vagueness. Professor Hogben (2) plays ducks and drakes with a battery which is able to write prescriptions, and, while disapproving of the everyday phrase put up with, is unwilling to look egregious up in the dictionary and see what it means. (3), if one takes an uncharitable attitude towards it, is simply meaningless: probably one could work out its intended meaning by reading the whole of the article in which it occurs. In (4), the writer knows more or less what he wants to say, but an accumulation of stale phrases chokes him like tea leaves blocking a sink. In (5), words and meaning have almost parted company. People who write in this manner usually have a general emotional meaning–they dislike one thing and want to express solidarity with another–but they are not interested in the detail of what they are saying. A scrupulous writer, in every sentence that he writes, will ask himself at least four questions, thus: What am I trying to say? What words will express it? What image or idiom will make it clearer? Is this image fresh enough to have an effect? And he will probably ask himself two more: Could I put it more shortly? Have I said anything that is avoidably ugly? But you are not obliged

Figure A.4. Politics and the English language by George Orwell (cont.).

to go to all this trouble. You can shirk it by simply throwing your mind open and letting the ready-made phrases come crowding in. They will construct your sentences for you–even think your thoughts for you, to a certain extent-and at need they will perform the important service of partially concealing your meaning even from yourself. It is at this point that the special connection between politics and the debasement of language becomes clear.

In our time it is broadly true that political writing is bad writing. Where it is not true, it will generally be found that the writer is some kind of rebel, expressing his private opinions and not a "party line." Orthodoxy, of whatever color, seems to demand a lifeless, imitative style. The political dialects to be found in pamphlets, leading articles, manifestoes, White Papers and the speeches of undersecretaries do, of course, vary from party to party, but they are all alike in that one almost never finds in them a fresh, vivid, home-made turn of speech. When one watches some tired hack on the platform mechanically repeating the familiar phrases–bestial atrocities, iron heel, bloodstained tyranny, free peoples of the world, stand shoulder to shoulder–one often has a curious feeling that one is not watching a live human being but some kind of dummy: a feeling which suddenly becomes stronger at moments when the light catches the speaker's spectacles and turns them into blank discs which seem to have no eyes behind them. And this is not altogether fanciful. A speaker who uses that kind of phraseology has gone some distance towards turning himself into a machine. The appropriate noises are coming out of his larynx, but his brain is not involved as it would be if he were choosing his words for himself. If the speech he is making is one that he is accustomed to make over and over again, he may be almost unconscious of what he is saying, as one is when one utters the responses in church. And this reduced state of consciousness, if not indispensable, is at any rate favorable to political conformity.

Figure A.4. Politics and the English language by George Orwell (cont.).

In our time, political speech and writing are largely the defense of the indefensible. Things like the continuance of British rule in India, the Russian purges and deportations, the dropping of the atom bombs on Japan, can indeed be defended, but only by arguments which are too brutal for most people to face, and which do not square with the professed aims of political parties. Thus political language has to consist largely of euphemism, question-begging and sheer cloudy vagueness. Defenseless villages are bombarded from the air, the inhabitants driven out into the countryside, the cattle machine-gunned, the huts set on fire with incendiary bullets: this is called pacification. Millions of peasants are robbed of their farms and sent trudging along the roads with no more than they can carry: this is called transfer of population or rectification of frontiers. People are imprisoned for years without trial, or shot in the back of the neck or sent to die of scurvy in Arctic lumber camps: this is called elimination of unreliable elements. Such phraseology is needed if one wants to name things without calling up mental pictures of them. Consider for instance some comfortable English professor defending Russian totalitarianism. He cannot say outright, "I believe in killing off your opponents when you can get good results by doing so." Probably, therefore, he will say something like this:

'While freely conceding that the Soviet regime exhibits certain features which the humanitarian may be inclined to deplore, we must, I think, agree that a certain curtailment of the right to political opposition is an unavoidable concomitant of transitional periods, and that the rigors which the Russian people have been called upon to undergo have been amply justified in the sphere of concrete achievement.'

The inflated style is itself a kind of euphemism. A mass of Latin words falls upon the facts like soft snow, blurring the outlines and covering up all the details. The great enemy of clear language is insincerity. When there is a gap between one's real and one's declared aims, one turns, as it were instinctively, to long words

Figure A.4. Politics and the English language by George Orwell (cont.).

and exhausted idioms, like a cuttlefish squirting out ink. In our age there is no such thing as "keeping out of politics." All issues are political issues, and politics itself is a mass of lies, evasions, folly, hatred and schizophrenia. When the general atmosphere is bad, language must suffer. I should expect to find–this is a guess which I have not sufficient knowledge to verify–that the German, Russian and Italian languages have all deteriorated in the last ten or fifteen years as a result of dictatorship.

But if thought corrupts language, language can also corrupt thought. A bad usage can spread by tradition and imitation, even among people who should and do know better. The debased language that I have been discussing is in some ways very convenient. Phrases like a not unjustifiable assumption, leaves much to be desired, would serve no good purpose, a consideration which we should do well to bear in mind, are a continuous temptation, a packet of aspirins always at one's elbow. Look back through this essay, and for certain you will find that I have again and again committed the very faults I am protesting against. By this morning's post I have received a pamphlet dealing with conditions in Germany. The author tells me that he "felt impelled" to write it. I open it at random, and here is almost the first sentence that I see: "[The Allies] have an opportunity not only of achieving a radical transformation of Germany's social and political structure in such a way as to avoid a nationalistic reaction in Germany itself, but at the same time of laying the foundations of a cooperative and unified Europe. " You see, he "feels impelled" to write–feels, presumably, that he has something new to say–and yet his words, like cavalry horses answering the bugle, group themselves automatically into the familiar dreary pattern. This invasion of one's mind by ready-made phrases (lay the foundations, achieve a radical transformation) can only be prevented if one is constantly on guard against them, and every such phrase anesthetizes a portion of one's brain.

Figure A.4. Politics and the English language by George Orwell (cont.).

I said earlier that the decadence of our language is probably curable. Those who deny this would argue, if they produced an argument at all, that language merely reflects existing social conditions, and that we cannot influence its development by any direct tinkering with words and constructions. So far as the general tone or spirit of a language goes, this may be true, but it is not true in detail. Silly words and expressions have often disappeared, not through any evolutionary process but owing to the conscious action of a minority. Two recent examples were explore every avenue and leave no stone unturned, which were killed by the jeers of a few journalists. There is a long list of fly-blown metaphors which could similarly be got rid of if enough people would interest themselves in the job; and it should also be possible to laugh the not un- formation out of existence,3 to reduce the amount of Latin and Greek in the average sentence, to drive out foreign phrases and strayed scientific words, and, in general, to make pretentiousness unfashionable. But all these are minor points. The defense of the English language implies more than this, and perhaps it is best to start by saying what it does not imply.

3. One can cure oneself of the not un- formation by memorizing this sentence: A not unblack dog was chasing a not unsmall rabbit across a not ungreen field.

To begin with, it has nothing to do with archaism, with the salvaging of obsolete words and turns of speech, or with the setting-up of a "standard-English" which must never be departed from. On the contrary, it is especially concerned with the scrapping of every word or idiom which has outworn its usefulness. It has nothing to do with correct grammar and syntax, which are of no importance so long as one makes one's meaning clear, or with the avoidance of Americanisms, or with having what is called a "good prose style"Ȯn the other hand it is not concerned with fake simplicity and the attempt to make written English colloquial. Nor does

Figure A.4. Politics and the English language by George Orwell (cont.).

it even imply in every case preferring the Saxon word to the Latin one, though it does imply using the fewest and shortest words that will cover one's meaning. What is above all needed is to let the meaning choose the word, and not the other way about. In prose, the worst thing one can do with words is to surrender them. When you think of a concrete object, you think wordlessly, and then, if you want to describe the thing you have been visualizing, you probably hunt about till you find the exact words that seem to fit it. When you think of something abstract you are more inclined to use words from the start, and unless you make a conscious effort to prevent it, the existing dialect will come rushing in and do the job for you, at the expense of blurring or even changing your meaning. Probably it is better to put off using words as long as possible and get one's meaning as clear as one can through pictures or sensations. Afterwards one can choose–not simply accept–the phrases that will best cover the meaning, and then switch round and decide what impressions one's words are likely to make on another person. This last effort of the mind cuts out all stale or mixed images, all prefabricated phrases, needless repetitions, and humbug and vagueness generally. But one can often be in doubt about the effect of a word or a phrase, and one needs rules that one can rely on when instinct fails. I think the following rules will cover most cases:
(i) Never use a metaphor, simile or other figure of speech which you are used to seeing in print.

(ii) Never use a long word where a short one will do.

(iii) If it is possible to cut a word out, always cut it out.

(iv) Never use the passive where you can use the active.

Figure A.4. Politics and the English language by George Orwell (cont.).

(v) Never use a foreign phrase, a scientific word or a jargon word if you can think of an everyday English equivalent.

(vi) Break any of these rules sooner than say anything barbarous.

These rules sound elementary, and so they are, but they demand a deep change of attitude in anyone who has grown used to writing in the style now fashionable. One could keep all of them and still write bad English, but one could not write the kind of stuff that I quoted in these five specimens at the beginning of this article.

I have not here been considering the literary use of language, but merely language as an instrument for expressing and not for concealing or preventing thought. Stuart Chase and others have come near to claiming that all abstract words are meaningless, and have used this as a pretext for advocating a kind of political quietism. Since you don't know what Fascism is, how can you struggle against Fascism? One need not swallow such absurdities as this, but one ought to recognize that the present political chaos is connected with the decay of language, and that one can probably bring about some improvement by starting at the verbal end. If you simplify your English, you are freed from the worst follies of orthodoxy. You cannot speak any of the necessary dialects, and when you make a stupid remark its stupidity will be obvious, even to yourself. Political language-and with variations this is true of all political parties, from Conservatives to Anarchists–is designed to make lies sound truthful and murder respectable. and to give an appearance of solidity to pure wind. One cannot change this all in a moment, but one can at least change one's own habits, and from time to time one can even, if one jeers loudly enough, send some worn-out and useless phrase–some jackboot, Achilles' heel, hotbed, melting pot, acid test, veritable inferno or other lump of verbal refuse–into the dustbin where it belongs.

Figure A.4. Politics and the English language by George Orwell (cont.).

A.5. Top 25 hedge fund managers earned $13bn in 2015 – more than some nations

The world's top 25 hedge fund managers earned $13bn last year – more than the entire economies of Namibia, the Bahamas or Nicaragua.

Kenneth Griffin, founder and chief executive of Citadel, and James Simons, founder and chairman of Renaissance Technologies, shared the top spot, taking home $1.7bn each – equivalent to the annual salaries of 112,000 people taking home the US federal minimum wage of $15,080.

The earnings of the best-performing hedge fund managers, published by Institutional Investor's Alpha magazine on Tuesday, dwarfs the pay of top Wall Street executives who have been under fire for their multimillion-dollar pay deals. The best paid banker last year was JPMorgan Chase CEO Jamie Dimon, who collected $27m.

The huge pay at the top comes despite a tumultuous year on Wall Street that has led many well-known hedge funds to lose billions of dollars and others to close down. Daniel Loeb, CEO of Third Point, a hedge fund that manages $17.5bn, has described market conditions as a "hedge fund killing field ".

Despite the challenges, Simons and Griffin managed to increase their earnings by

Figure A.5. Top 25 hedge fund managers earned $13bn in 2015 – more than some nations.

$500m and $400m, respectively, compared with last year.

Both men have poured a lot of money into the presidential race, but both backed Republicans who dropped out. Griffin, who is the richest man in Illinois with a $7.5bn fortune according to Forbes, has donated more than $3m into the failed campaigns of Marco Rubio, Jeb Bush and Scott Walker.

Griffin, 47, who started from his dorm at Harvard University, was the biggest single donor to Rahm Emanuel's successful campaign for a second term as mayor of Chicago.

He has rarely spoken about his political inclinations, but in 2012 he described himself as a "Reagan Republican" and said he thought the rich had "insufficient influence" on the political process. When Emanuel announced the closure of 50 schools, Griffin said he should have closed 125.

Griffin recently spent $500m buying Jackson Pollock's Number 17A and Willem de Kooning's Interchanged from the entertainment mogul David Geffen. He has loaned the paintings to the Art Institute of Chicago.

Simons, a string theory expert and former cold war codebreaker, has made an estimated $15.5bn from Renaissance Technologies the mathematics-driven "quant" hedge fund he set up 34 years ago.

The fund, which is run from the tiny Long Island village of Setauket where Simons owns a huge beachfront compound, has donated $13m to Cruz's failed campaign. With Cruz out of the race, Renaissance has switched donations to Hillary Clinton, with more than $2m donated so far. Euclidean Capital, Simon's family office, has donated more than $7m to Clinton.

Figure A.5. Top 25 hedge fund managers earned $13bn in 2015 – more than some nations (cont.).

Simons, 78, who retired as CEO of Renaissance in 2009, is the 50th richest person in the world, according to Forbes. His earnings last year were so large that if he were a country it would rate as the world's 178th most productive nation, according to the World Bank's GDP rankings.

He has donated millions of dollars to maths and science education via the Simons Foundation he set up in 1994.

No woman has yet made it into the top 25 of the hedge fund highest-paid list, which has been running for 15 years. Hedge fund managers typically get paid based on a structure known as "two and 20", in which they collect a 2% fee on the assets they manage and earn 20% of the profits they make for investors.

Figure A.5. Top 25 hedge fund managers earned $13bn in 2015 – more than some nations (cont.).

A.6. Sats tests will harm next generation of writers, says Society of Authors

Children's authors are warning that the "restrictive" way children in England are being taught writing in school will affect the next generation of novelists, biographers and poets.

In a statement released by members of the Society of Authors who write for children

Figure A.6. Sats tests will harm next generation of writers, says Society of Authors.

and for education, they condemn current government policy on the teaching of writing and grammar. They say the government has intervened too far and that "the resultant teaching no longer reflects what writing really does".

As year 6 children sit their Sats tests this week – including spelling, punctuation and grammar – the authors say that when the Department for Education introduces new terminology for grammatical structure, such as "fronted adverbs", and insists that exclamation marks can only end sentences starting with "what" or "how", it risks "alienating, confusing and demoralising children with restrictions on language just at the time when they need to be excited by the possibilities".

The statement calls on the government to "allow the current generation of schoolchildren in England to enjoy language, to be empowered by their skill in it, and not to become tangled in rules which have no application outside the narrow confines of a National Test".

"Amongst these children must be the next generation of novelists, screenwriters, biographers, poets and science writers. We need our children to become fluent, eager and expressive writers, able to persuade, entrance and uplift with language, able to create empathy and delight in their readers. We cannot risk destroying their enjoyment, confidence and power at such an early age," the authors say.

The Carnegie medal-winner David Almond, author of Skellig and a former teacher, added that children "instinctively know [that language] is a fluid, flexible, beautiful thing", and that they "learn how to talk, to sing, to converse by falling in love with language, by delighting in their own skills, by sharing and exploring those skills with others". Current government policy, Almond said, "interferes with this process".

Figure A.6. Sats tests will harm next generation of writers, says Society of Authors (cont.).

"We do our children great harm by insisting too early that they analyse and explain exactly what they are doing. Such an approach is deeply pessimistic," he said. "Why do we not trust, celebrate and encourage the natural human ability to explore, celebrate, enjoy and control language? Why do we want to tell our children that they are wrong and that they fail? Why do government ministers think they know more than teachers who have devoted their lives to the education of the nation's children?"

Author Anne Rooney, chair of the Society of Authors' educational writers group committee, attacked in particular the new rule on exclamation marks, saying that if children come across exclamation marks in books, they will wonder why the rules they have been taught don't match what they see in practice. "It's not as though exclamation marks are only safe in the hands of grown-ups. It isn't like not letting them drive or drink alcohol or join the army – all things they can do when they are older but are against the rules in primary school. No one is going to be hurt by a sharp exclamation mark," she writes.

Nicola Morgan, chair of the Society's children's writers and illustrators group committee, said that the government's "desperation" to measure risks "throwing everything else out: structure and style, clarity and beauty. And love of language. While teaching some bonkers 'rules' along the way."

The new statement comes the week after Sats tests for seven-year-olds were boycotted by some families, while a letter to the Guardian at the end of last month, signed by writers including Philip Pullman and Michael Rosen, called for "2016 to be the final year of primary assessment in its current form" and for the Sats "to go".

Figure A.6. Sats tests will harm next generation of writers, says Society of Authors (cont.).