Across-Time Comparative Summarization of News Articles

Yijun Duan
Graduate School of Informatics, Kyoto University
Kyoto, Japan
yijun@dl.kuis.kyoto-u.ac.jp

Adam Jatowt
Graduate School of Informatics, Kyoto University
Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Comparative summarization is an effective strategy to discover important similarities and differences in collections of documents biased to users' interests. A natural method of this task is to find important and corresponding content. In this paper, we propose a novel research task of automatic query-based across-time summarization in news archives as well as we introduce an effective method to solve this task. The proposed model first learns an orthogonal transformation between temporally distant news collections. Then, it generates a set of corresponding sentence pairs based on a concise integer linear programming framework. We experimentally demonstrate the effectiveness of our method on the New York Times Annotated Corpus.

CCS CONCEPTS

• **Information systems → Summarization**; • **Computing methodologies → Information extraction**;

KEYWORDS

Query-based comparative summarization; embeddings alignment; integer linear programming, news archives

ACM Reference Format:
Yijun Duan and Adam Jatowt. 2019. Across-Time Comparative Summarization of News Articles. In *Proceedings of ACM WSDM conference (WSDM'19).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Nowadays, news is one of the most important channels for acquiring high-quality information regarding our society. However, with the rapid growth of Web, more and more news articles are available causing information overload. News summarization can help to combat this problem by distilling the most important information from large amount of news articles for users.

Sometimes, users wish to compare two collections of news, which can be from distant times, biased to a given specific query (that conveys the interests of users) to discover commonalities and differences between them. For example, users may be interested in comparison of politicians in 1980s with those in 2010s. They would like to read corresponding and important news related to the query "politician" from these two periods. Corresponding news are news containing comparable aspects. For example, "president Trump visited China in 2017" and "president Nixon visited China in 1972" are corresponding news because Nixon and Trump are temporally corresponding entities. Note that such corresponding relationship is difficult to be identified by average users since they tend to possess less knowledge about the past than about the present. Users however can benefit from this kind of information for needs such as analyzing trends, gaining insights about similar situations in the past, making better decisions and so on.

The input collections of news for the comparative summary can be however very diverse and may cover several latent subgroups (e.g., many diverse news are published every year that relate to politicians). Thus instead of a single output news pair, a set of pairs that represent latent subgroups within the input collections should be returned as a summary. In the previously mentioned example, apart from events related to U.S. presidents, important news about presidents of other major countries should also be covered in the resulting summary.

In this paper, we propose a novel task of *query-based across-time comparative news summarization* that generates query-dependent contrastive summary between two news collections from different time periods (typically, one representing the present time and the other being some period in the past). The output summary is in the form of pairs of corresponding news contents. We set up four objectives for this task. The first one is *relevance*. It requires the news pairs in the formed summary be relevant to the query given by the user. The second one is *correspondence*. News in the same pair should demonstrate good correspondence and comparability towards alleviating the context gap between the temporally-distant input collections. The third one is *saliency*. Summary news should be important capturing the majority of information in a news article collection. The news with minor significance should not be included in the summary. The last goal is *diversity*. The information overlap between the selected news pairs should be as minimal as possible due to the length restriction of summary. The summary should thus cover diverse aspects of information.

The problem of across-time comparative news summarization is not trivial resulting from the following reasons: (1) To measure sentence correspondence is a difficult task. The general context of the two input news collections which originate from different time periods may be fairly different. For example, there is low overlap between the contexts of newspaper articles across time gaps spanning more than 20 years. Intuitively, the correspondence of news in different contexts cannot be computed properly without a solid understanding of the connection (analogies) between their contexts. Moreover, it is difficult to collect training data for learning such connections. (2) Constructing an optimal summary in view of all the above-listed criteria (relevance, correspondence, saliency,

WSDM'19, February 2019, Melbourne, Australia

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

diversity) is a challenging problem since the selected news pairs should satisfy all the aforementioned criteria in an optimal way.

For solving these challenges, we propose a novel summarization system to address the proposed task. First of all, we formulate the measurement of sentence correspondence by finding corresponding terms and aligning different vector spaces. We first adopt the distributed vector representation to represent the context vectors of words; then we learn linear and orthogonal transformations between two vector spaces of input collections for learning word correspondence. Furthermore, we measure the correspondence between two candidate news as the minimal amount of distance that the words of one news need to "travel" to match the words of another news, suggested by the commonly-used Word Mover's Distance (WMD) algorithm [20]. Secondly, inspired by the popular Affinity Propagation algorithm [9], we propose a concise joint integer linear programming framework which detects diverse and representative news (which we call exemplars) and at the same time generates correspondent pairs from the detected exemplars. Based on this formulation, exactly optimal solution can be obtained.

To sum up, we make the following contributions:

(1) We introduce a new research problem of automatically generating a summary of corresponding news from two temporally-distant collections of articles based on a user query. Our proposal is essentially a novel way to accessing information from news archives as we summarize past articles through "present lens" by discovering the commonalities between the query-related content in the past and one in the present.
(2) We develop a novel summarization system to address this task relying on an effective news correspondence measurement and a concise integer linear programming framework.
(3) Finally, we perform extensive experiments on the New York Times Annotated Corpus, which prove the effectiveness of our approach.

The remainder of this paper is structured as follows. We first formally define the research problem in Section 2. We survey the related work in Section 3. The across-time news correspondence measurement and ILP formulation for summarization are presented in Section 4. We describe the experiments on transformation and summarization in Section 5 and Section 6, respectively. Finally, the last section concludes the paper and outlines the future work.

2 PROBLEM DEFINITION

Formally, let D_A and D_B denote two sets of newspaper articles, where D_A and D_B have different publication time periods T_A and T_B, respectively ($T_A \cap T_B = \emptyset$ and, typically T_A represents the present time period while T_B represents some period in the past). In this paper we focus on the extractive summarization methods which extract the summary sentences directly from the input documents. Then given a specific user query q, the task is to discover m relevant sentence pairs $P = [p_1, p_2, ..., p_m]$ to form a concise summary conveying the most important comparisons, where $p_i = (s_i^A, s_i^B)$ and s_i^A and s_i^B are sentences extracted from D_A and D_B, respectively. The pairs included in the summary should have good quality, i.e., sentences within the same pair should deliver comparable aspects. Moreover, the selected pairs should satisfy the aforementioned criteria of relevance, saliency and diversity.

3 RELATED WORK

To the best of our knowledge, the research task of query-based comparative news summarization across time has not been proposed neither approached so far. Our work is nevertheless connected to several research areas: *query-focused summarization*, *comparative summarization* and *embeddings alignment* that are surveyed below.

3.1 Comparative Summarization and Query-focused Summarization

The task of automatic document summarization aims at generating short summaries for long documents. In recent years there have been multiple studies addressing different scenarios for document summarization beyond the traditional generic task [41]. Among those, *comparative summarization* focuses on generating summaries from multiple comparative aspects. Recent approaches to this task rely on a multivariate normal generative model [39], a hierarchical non-parametric Bayesian model [31], differential topic models [13] and integer linear programming models [14].

Query-focused summarization addresses another goal to require the summarization process be guided by a pre-specified query term which specifies the information need of the user. Until now, typical query-focused systems simply had no special treatments in comparison to generic systems other than incorporating literal similarity between documents and query. Recent approaches to this task include a sentence compression models [4], graph-based models [21], learning to rank models [34], and regression models [28].

To the best of our knowledge, we are the first to focus on the task of providing users with a summary containing temporally corresponding sentences based on the search query. We also propose a novel and concise integer linear programming framework for this summarization task.

3.2 Temporal Analog Detection and Embeddings Alignment

A part of our system approaches the task of identifying temporally corresponding sentences across different times. The related works to this subtask include detecting temporal changes in word meaning [15, 18, 19, 25] and computing term similarity across time [1, 16, 17, 37]. Our work is however different as we consider structural correspondence between different time periods. In this study we represent terms using the distributed vector representation [26]. Thus the problem of connecting news articles' context across different time periods can be approached by aligning pre-trained word embeddings in different time periods. Mikolov *et al.* proposed a linear transformation aligning bi-lingual word vectors for automatic text translation such as translation from Spanish to English [27]. Faruqui *et al.* obtained bi-lingual word vectors using CCA [8]. More recently, Xing *et al.* argued that the linear matrix adopted by Mikolov *et al.* should be orthogonal [40]. Similar suggestion has been given by Samuel *et al.* [35]. Besides linear models, non-linear models such as "deep CCA" has also been introduced for the task of mapping multi-lingual word embeddings [24]. In this study we adopt the orthogonal transformation for computing across-time term correspondence due to its high accuracy and efficiency. We then develop a sentence-level correspondence metric based on the widely-used distance function WMD [20].

The most similar work to ours is one focusing on search of corresponding terms across time [42]. Our research problem is however more difficult. Not only we discover corresponding sentences instead of words, but we also utilize other information such as relevance and saliency of the extracted sentences.

4 PROPOSED METHOD

In this section, we first describe our approach for satisfying four objectives of good candidate sentences including *relevance*, *saliency*, *diversity* and *correspondence*. We then introduce the integer linear programming framework proposed for extracting summary sentences considering all the described objectives.

4.1 Relevance

We begin by introducing the method used for discovering sentences relevant to a specific user query q in news archives. Generally, for query-based summarization, the query information is considered via computations of semantic overlap between the query and each sentence. We deploy the popular ranking model $BM25$ [32] to rank top relevant sentences (i.e., top-100) according to their similarity to q in archival document collections D_A and D_B as an initial run (see Sec. 6). We denote the retrieved candidate sentences (i.e., top-n sentences) as D'_A and D'_B.

4.2 Saliency and Diversity

Summary sentences should be representative sentences which capture salient information in a news archive. In addition, the information overlap between summary sentences in different pairs should be as minimal as possible for avoiding redundancy. However, the prior here is that the ratio of salient candidate sentences over trivial candidate sentences is very low given the summary length limit, and the input news archives can be very diverse as well as may cover multiple latent subgroups. We thus call a sentence as *exemplar* if it represents a group of sentences based on some measure of similarity (e.g., word mover's distance). Intuitively, if the selected exemplars concisely represent the entire set of sentences, the saliency and diversity will naturally arise. Note that whether sentence s_i is selected as an exemplar or not, and whether sentence s_j is represented by s_i or not are latent information.

Formally, let $G(s_i)$ denote the set of sentences which vote for s_i as their exemplar (i.e., are represented by that exemplar). The saliency of s_i denoted as $Imp(s_i)$ is computed using Word Mover's Distance (WMD) algorithm [20]:

$$Imp(s_i, G(s_i)) = \sum_{s_j \in G(s_i)} Sim_{wmd}(s_i, s_j) \tag{1}$$

4.3 Correspondence

In this section, we describe the method for measuring temporal correspondence between a sentence s_A in set D'_A and a sentence s_B in the other set D'_B. Intuitively, if s_A and s_B correspond to each other, then s_A and s_B contain correspondent and comparable aspects. For instance, <*iPod*, *Walkman*> could be regarded as a pair of correspondent aspects based on the observation that *Walkman* played the role of a popular portable music player 30 years ago same as *iPod* does nowadays. The key difficulty comes from the fact that there is low overlap between terms' contexts across time (e.g., the set of top co-occurring words with *iPod* in documents published in 2010s has typically little overlap with the set of top co-occurring words with *walkman* that are extracted from documents in 1980s). Thus our task is then to build the connection between semantic spaces of D_A and D_B.

Let transformation matrix W map the words from D_A into D_B, and transformation matrix Q map the words in D_B back into D_A. Let a and b be normalized word vectors from the news document collections D_A and D_B, respectively. The correspondence between words a and b can be evaluated as the similarity between vectors b and Wa, i.e., $Corr(a, b) = b^T Wa$. However we could also form this correspondence as $Corr'(a, b) = a^T Qb$.

Theorem 4.1. *The linear transformations W and Q between spaces D_A and D_B should be orthogonal.*

Proof. To be self-consistent, we require $Corr(a, b) = Corr'(a, b)$, thus $b^T Wa = (b^T Wa)^T = a^T W^T b = a^T Qb$, and therefore $Q = W^T$. Furthermore, when we map a term from D_A into D_B, we should be able to map it back into D_A and obtain the original vector, hence, $a = Q(Wa) = W^T(Wa)$. This expression should hold for any term in D_A and we conclude that the transformation W should be an orthogonal matrix satisfying $W^T W = I$ where I denotes the identity matrix. Thus transformations W and its transpose Q are orthogonal. □

This observation has also been reported in [35, 40] for the purpose of bi-lingual text translation. Note that orthogonal transformation preserves vector norms, so given normalized vectors a and b, $Corr(a, b) = b^T Wa = |b||Wa|cos < b, Wa >= cos < b, Wa >= Sim_{cosine}(a, b)$.

However, the following challenge is that the training term pairs for learning the mapping W are difficult to obtain. We adopt here a trick proposed by [42] for preparing enough training data. Namely, we use so-called Common Frequent Terms (CFT) as the training term pairs. CFT are very frequent terms in both compared document collections (e.g. man, woman, sky, water). Such frequent terms tend to change their meanings only to a small extent across time. The phenomenon that words which are intensively used in everyday life evolve more slowly has been reported in several languages including English, Spanish, Russian and Greek [12, 22, 29][1]. Given L pairs composed of normalized vectors of CFTs trained in both document collections [(a_1,b_1), (a_2,b_2), ..., (a_L,b_L)], we should learn the transformation W by maximizing the accumulated cosine similarity of CFT pairs (we utilize the top 7% (25k words) of Common Frequent Terms to train the transformation matrix. In Sec. 6. we discuss the transformation performance w.r.t. different numbers of used CFTs),

$$\max_W \sum_{i=1}^{L} b_i^T W a_i, s.t. W^T W = I \tag{2}$$

To infer the orthogonal transformation W from pairs of CFTs $\{a_i, b_i\}_{i=1}^L$, we state the following theorem.

Theorem 4.2. *Let A and B denote two matrices, such that the i^{th} row of (A, B) corresponds to pair of vectors (a_i^T,b_i^T). By computing*

[1] Note that even if certain CFTs do not retain the same semantics across time, the results should not deteriorate significantly when the number of used CFTs is sufficiently high.

the SVD of $M = A^T B = U\Sigma V^T$*, the optimized transformation matrix* W^* *satisfies*

$$W^* = U \cdot V^T \tag{3}$$

Proof. Maximizing $\sum_{i=1}^{L} b_i^T W a_i$ equals to maximizing $tr(BWA^T)$ $= tr(A^T BW) = tr(U\Sigma V^T W) = tr(\Sigma V^T WU)$. Let $Z = V^T WU$, for Z is orthogonal (being the product of orthogonal matrices), thus $\sum_j Z_{i,j}^2 = 1$ and $Z_{i,i} \leq 1$. Then $tr(\Sigma V^T WU) = tr(\Sigma Z) = \sum_i \Sigma_{i,i} Z_{i,i} \leq \sum_i \Sigma_{i,i}$. The last inequality holds because $\Sigma_{i,i} \geq 0$ (given Σ is obtained by SVD). Then the objective can achieve the maximum if $Z_{i,i} = 1$ which implies $Z = I$. So an optimal W^* is UV^T. □

The obtained orthogonal transformation W^* allows for learning correspondence on the level of terms. Based on it, we measure the correspondence between two sentences, $s_A = [w_1^A, w_2^A, ..., w_m^A]$ and $s_B = [w_1^B, w_2^B, ..., w_n^B]$ as the maximum amount of cosine similarity that the words of s_B can obtain to match the words of $W^* \cdot s_A = [W^* \cdot w_1^A, W^* \cdot w_2^A, ..., W^* \cdot w_n^A]$, as suggested by the commonly-used Word Mover's Distance (WMD) algorithm [20].

$$Corr(s_A, s_B) = Sim_{wmd}(W^* \cdot s_A, s_B) \tag{4}$$

4.4 Joint ILP Formulation (J-ILP) for Summarization

In this section, we describe our approach for generating temporally corresponding sentence pairs as summary. Given two sets of initially retrieved sentences D'_A and D'_B associated with time periods T_A and T_B, respectively, the output summary is composed of m pairs of comparable sentences $[p_1, p_2, ..., p_m]$, where each pair contains a sentence from D'_A and from D'_B.

Inspired by the popular Affinity Propagation algorithm [9] for exemplar-based clustering, we formulate our task as a process of identifying a subset of corresponding exemplar sentence pairs which concisely convey the most import comparisons between input datasets. We propose a Joint Integer Linear Programming (J-ILP) formulation for discovering comparable sentences, and we use the *bound-and-branch* method to obtain the optimal solution.

More explicitly, we formulate the task as a process of selecting a subset of k_A and k_B exemplars for each news collection, respectively, and ranking m exemplar pairs based on the identified exemplars. Each non-exemplar sentence is assigned to an exemplar sentence based on their WMD distance, and each exemplar e represents a subgroup comprised of all non-exemplar sentences that are assigned to e. On the one hand, we wish to maximize the overall saliency of selected exemplars w.r.t. their representing subgroups. On the other hand, we expect to maximize the overall correspondence of the top m sentence pairs across time, where each pair consists of two exemplars from different document collections.

We next introduce notations used in our method. Let s_i^A denote the ith sentence in D'_A. $M_A = [m_{ij}]^A$ is a $n \times n$ binary square matrix such that n is the number of initially retrieved sentences in D'_A. m_{ii}^A indicates whether sentence s_i^A is selected as an exemplar or not, and $m_{ij:i\neq j}^A$ represents whether sentence s_i^A votes for sentence s_j^A as its exemplar. Similar to M_A, the $n \times n$ binary square matrix M_B indicates how sentences belonging to D'_B choose their exemplars. m_{ii}^B indicates whether sentence s_i^B is selected as an exemplar or not, and $m_{ij:i\neq j}^B$ represents whether sentence s_i^B votes for sentence s_j^B as its exemplar. Different from M_A and M_B, $M_T = [m_{ij}]^T$ is a $n \times n$ binary matrix whose entry m_{ij}^T denotes whether sentences s_i^A and s_j^B are paired together as the final result. Then the following ILP problem is designed for the task of selecting k_A and k_B exemplars for each document collection, respectively, and for ranking m sentence pairs (the formulas are described in the next paragraph):

$$\begin{aligned} max \lambda \cdot m \cdot \left[I'(M_A) + I'(M_B)\right] \\ + (1-\lambda) \cdot (k_A + k_B) \cdot C'(M_T) \end{aligned} \tag{5}$$

$$I'(M_X) = \sum_{i=1}^{n} m_{ii}^X \cdot Imp(s_i^X, G(s_i^X)), X \in \{A, B\} \tag{6}$$

$$C'(M_T) = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}^T \cdot Corr(s_i^A, s_j^B) \tag{7}$$

$$\begin{aligned} G(s_i^X) = \left\{ s_j^X | m_{ji}^X = 1, j \in \{1, ..., n\} \right\}, \\ i \in \{1, ..., n\}, X \in \{A, B\} \end{aligned} \tag{8}$$

$$\begin{aligned} s.t. \quad m_{ij}^X \in \{0, 1\}, i \in \{1, ..., n\}, \\ j \in \{1, ..., n\}, X \in \{A, B, T\} \end{aligned} \tag{9}$$

$$\sum_{i=1}^{n} m_{ii}^X = k_X, X \in \{A, B\} \tag{10}$$

$$\sum_{j=1}^{n} m_{ij}^X = 1, i \in \{1, ..., n\}, X \in \{A, B\} \tag{11}$$

$$\begin{aligned} m_{jj}^X - m_{ij}^X \geq 0, i \in \{1, ..., n\}, \\ j \in \{1, ..., n\}, X \in \{A, B\} \end{aligned} \tag{12}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}^T = m \tag{13}$$

$$m_{ii}^A - m_{ij}^T \geq 0, i \in \{1, ..., n\}, j \in \{1, ..., n\} \tag{14}$$

$$m_{jj}^B - m_{ij}^T \geq 0, i \in \{1, ..., n\}, j \in \{1, ..., n\} \tag{15}$$

$$\sum_{j=1}^{n} m_{ij}^T \leq 1, i \in \{1, ..., n\} \tag{16}$$

$$\sum_{i=1}^{n} m_{ij}^T \leq 1, j \in \{1, ..., n\} \tag{17}$$

We now explain the meaning of the above formulas. First, Eq. (10) forces that k_A and k_B exemplars are identified for both sets D'_A and D'_B, respectively, and Eq. (13) guarantees that m sentence pairs are selected as the final result. The restriction given by Eq. (11) means that each sentence must choose only one exemplar. Eq. (12) enforces that if one sentence s_j^X is voted by at least one other sentence, then it must be an exemplar (i.e., $m_{jj}^X = 1$). The constraints given by Eq. (14) and (15) jointly guarantee that if a sentence is selected in any corresponding sentence pair (i.e., $m_{ij}^T = 1$), then it must be an exemplar in its own subgroup (i.e., $m_{ii}^A = 1$ and $m_{jj}^B = 1$).

Restricted by Eq. (16) and Eq. (17), each selected exemplar in the result is only allowed to appear once to avoid redundancy.

$I'(M_X)$ depicts the overall saliency of selected exemplars in both sets D'_A and D'_B, respectively. $G(s_i^X)$ denotes the representing subgroup for sentence s_i^X (if s_i^X is not chosen as an exemplar, its representing subgroup will be null). $C'(M_T)$ depicts the overall correspondence of the generated sentence pairs. In view of the fact that there are $(k_A + k_B)$ values (each value is in [0,1]) in the saliency part $I'(M_A) + I'(M_B)$, and there are m values (each value is in [0,1]) in the correspondence part $C'(M_T)$, we add the coefficients m and $(k_A + k_B)$ in the objective function to avoid suffering from skewness problem. Finally, the parameter λ [2] is used to balance the weight of the two parts.

Our proposed J-ILP formulation guarantees to achieve the optimal solution by using *bound-and-branch method*. We use the Gurobi solver [10] for solving the proposed J-ILP framework.

5 TRANSFORMATION EFFECTIVENESS

5.1 Datasets

We perform the experiments on the New York Times Annotated Corpus [33]. This corpus is a collection of 1.8 million articles published by the New York Times between January 01, 1987 and June 19, 2007 and has been frequently used to evaluate different researches that focus on temporal information processing or extraction in document archives [2]. For the experiments, we first divide the corpus into four parts according to article publication dates: [1987, 1991],[1992, 1996], [1997, 2001] and [2002, 2007]. The vocabulary size of each time period is around 300k. We then focus on comparing the pair of time periods which are separated by the longest time gap, [1987, 1991] (denoted as T_A) and [2002, 2007] (denoted as T_B). We assume here that the more farther the two time periods are apart, the stronger is the context change, which increases the difficulty of finding corresponding news. We obtain the distributed vector representations for both time periods T_A and T_B by training the Skip-gram model using the gensim Python library [30]. The number of dimensions of word vectors is experimentally set to 200.

5.2 Experimental Settings

5.2.1 Analyzed Methods. We first compare the performance of transformation models discussed in Sec. 4.3. The linear transformation model [27, 42] (denoted as *LT*) and orthogonal transformation model [35, 40] (denoted as *OT*) are investigated. Besides, we also test the method which directly compares the vectors in different time periods without performing any transformation (denoted as *Non-Tran*). We adopt the same parameter setting as in [42].

5.2.2 Test sets. To prove the transformation effectiveness of our approach, we focus on the task of searching for temporal analogs by transforming the term representations. We utilize the test sets containing queries in the base time [2002, 2007] and their analogs in target time [1987, 1991] which were used in [42]. The examples of the test queries and their temporal analogs are shown in Tab. 1 where q denotes the input term and *ta* is the correct temporal analog. In total, there are 95 pairs of terms (query and its analog) resulting from 54 input query terms for matching [2002, 2007] and [1987, 1991].

5.2.3 Evaluation metrics. The *Mean Reciprocal Rank (MRR)* is used for evaluating the search results for each transformation model, and is computed as follows:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{18}$$

where $rank_i$ is the rank of a correct temporal analog at the i-th test, and N is the number of test pairs. In addition, precisions @1, @5, @10 and @20 are also reported. Those metrics refer to the rates of tests in which the correct temporal analog was included in the top 1, 5, 10 and 20 results, respectively. All the values of used metrics fall into [0,1]. The higher the values are, the more effectively a given transformation model works.

5.3 Experimental Results

We show the detailed results for a few examples in Tab. 1, while Tab. 2 shows the scores for all the methods averaged on all the tested queries. We first notice that the performance is extremely poor without transforming the contexts of queries. The correct answers in *Non-Tran* approach are usually found at ranks > 1k. On the other hand, both mapping matrices *OT* and *LT* are quite helpful since they exhibit significantly better effectiveness in terms of all the metrics. This observation suggests little overlap in the contexts of news articles separated by longer time gaps, and that the task of temporal analog identification is quite difficult.

Moreover, a closer look at Tab. 2 reveals that regardless of the type of evaluation metric, an orthogonal subjection improves the performance of the linear model. Specifically, *OT* improves *LT* model by 8.0% when measured using the main metric MRR. This is also consistent with previous reports in aligning bi-lingual word vectors in language translation [35]. The plausible reason is that the orthogonal transformation allows for a self-consistent mapping between semantic spaces. In that case, when we map a source word into the target time, we are able to map it back into the source time and obtain its original vector. Such guarantee leads to a superior robustness of transformation and across-time analog detection.

6 SUMMARIZATION EFFECTIVENESS

6.1 Datasets

We use the same news article datasets as described in Sec. 5.1.

6.2 Test Sets

6.2.1 Test queries. In this study, we randomly select query terms from among Common Frequent Terms (CFT) shared by news articles in both T_A and T_B. We utilize the top 7% (25k words) of Common Frequent Terms to generate queries. The randomly chosen test queries cover both general topics (e.g. politics) and specific entities (e.g. Japan). The types of entities include locations, organizations and objects. In total, there are 30 queries used as shown in Tab. 3.

6.2.2 Initial retrieval. Given each query q, we apply the widely-used ranking model *BM25* [32] to rank the top-100 news articles according to their relevance to q in news archival collections of

[2] We experimentally set the value of λ to be 0.4 in Sec. 6.

Table 1: Example results where q is the input term in [2002, 2007] and ta is the matching temporal analog in [1987, 1991]. The numbers are the ranks of the correct temporal analog in the results ranked by each method. Ranks lower than 1000 are represented as 1000+.

q	ta	Non-Tran	LT	OT	q	ta	Non-Tran	LT	OT
mp3	cassette	1000+	**3**	29	sepp_blatter	joao_havelange	1000+	2	**1**
pixar	disney	1000+	**1**	3	ipod	walkman	1000+	12	**9**
linux	unix	1000+	**1**	**1**	email	fax	1000+	**17**	18
vladimir_putin	mikhail_gorbachev	1000+	45	**5**	email	letter	1000+	**1**	**1**
slovakia	czechoslovakia	1000+	19	**4**	berlin	bonn	1000+	3	**1**

Table 2: Results of searching from present to past (present: 2002-2007; past: 1987-1991).

Method	MRR	P@1	P@5	P@10	P@20
Non-Tran	0.015	0.000	0.024	0.024	0.024
LT	0.348	0.238	0.452	0.476	0.595
OT	**0.376**	**0.285**	**0.500**	**0.548**	**0.643**

Table 3: Summary of test queries.

Query Term
new_york, american, president, united_states, japan, war, washington, women, politics, university, military, economy, china, europe, finance, health, campaign, player, new_jersey, british, industry, republican, education, investment, corporation, california, soviet_union, democrat, baseball, technology

both T_A and T_B, respectively. *BM25* is deployed based on the default setting of Apache Solr search engine (k1=1.2, b=0.75).

6.2.3 Data annotation. In order to assess the quality of generated summaries by different test methods, three human judges who are not authors of this paper manually annotated the news articles. After reading the content of two initially retrieved document sets for each query, the annotators were asked to conduct the following three data annotation tasks: (1) *Task 1.* The first task was to highlight all the salient sentences in both sets to form the salient sentences set; (2) *Task 2.* The second task was to select all the temporally correspondent sentence pairs containing sentences from different sets. There was no limit on the number of sentences nor sentence pairs in the first two tasks. (3) *Task 3.* The third task was to write up to 300-words long *reference summary* of the highlighted text in the first two tasks, that will help in grasping the important and corresponding content of temporally distant news collections.

6.3 Analyzed Methods

We implemented three types of summarization models as analyzed methods in order to compare the summaries generated by them with the human-created reference summaries.

Type 1. We first test the performance of our proposed joint ILP formulation (J-ILP) based on different transformation models (*OT*, *LT* and *Non-Tran*). We experimentally set the value of weighting factor λ in Sec. 6 to be 0.4.

Type 2. We then make comparisons with 3 commonly used multi-document summarization techniques: *LexRank* [7], *LSA* [36], and *KLSUM* [11]. The frequently used *MMR* [3] method is adopted for all the three methods to diversify the generated summaries (the control parameter in MMR method was set to be 0.3, following the previous work [6]).

Type 3. The third type of strategy performs comparative summarization on the compared datasets, aiming to generate summaries that consider discriminative characteristics. Three state-of-the-art comparative summarization approaches which rely on a multivariate normal generative model [39] (denoted as *DSS*), a mutually-reinforced random walk model [5] (denoted as *MRRW*) and a comparative linear programming model [14] (denoted as *LPCM*) respectively are used for evaluation.

After the summary have been constructed by the above *Type 2* and *Type 3* methods, we build the sentence pairs which have the maximal correspondence based on summary sentences as follows.

$$\mathrm{P} \equiv \ \operatorname{argmax} \sum_{i=1}^{m} Corr(s_i^A, s_i^B) \tag{19}$$

where $P = [p_1, p_2, ..., p_m]$ are expected comparables, and $p_i = (s_i^A, s_i^B)$. s_i^A and s_i^B are summary sentences generated from the compared document sets.

6.4 Evaluation Metrics

We evaluate all the models with the following measures:

ROUGE-1.5.5 toolkit [23]. The ROUGE is a widely used metric which has been officially adopted by DUC for automatic summarization evaluation. It measures summary quality by counting overlapping units between the candidate summary and the reference summary (i.e., the summary created by judges in Data Annotation Task 3) [23]. In the experiment, we report the f-measure values of ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W and ROUGE-SU, which are based on overlapping unigrams, bigrams, longest common subsequence (LCS), weighted LCS and skip-bigram plus unigram respectively. The higher the ROUGE scores are, the more similar the machine-generated summary and the reference summary are, and thus the more effective the summarization approach is.

Precision. The ROUGE measures mainly reflect the recall. We further examine the rate of summary sentences and pairs included in the human-labeled important sentence set (created in Data Annotation Task 1) and sentence pair set (created in Data Annotation Task 2), respectively, as follows:

$$Precision_S = \frac{\{summary\_sentences\} \cap \{labeled\_sentences\}}{\{summary\_sentences\}} \tag{20}$$

$$Precision_P = \frac{\{summary\_sentence\_pairs\} \cap \{labeled\_sentence\_pairs\}}{\{summary\_sentence\_pairs\}} \tag{21}$$

Table 4: Overall performance comparison by each method using ROUGE.

Type	System	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
Proposed Methods	J-ILP (OT)	**0.485**	0.218	**0.447**	**0.195**	**0.148**
	J-ILP (LT)	0.478	**0.234**	0.423	0.162	0.128
	J-ILP (Non-Tran)	0.171	0.061	0.098	0.040	0.017
Multi-Document Summarization Methods	LexRank	0.260	0.138	0.186	0.089	0.064
	LSA	0.313	0.144	0.242	0.096	0.073
	KLSUM	0.235	0.048	0.169	0.085	0.021
Comparative Summarization Methods	DSS	0.179	0.075	0.151	0.074	0.057
	MRRW	0.346	0.180	0.283	0.103	0.077
	LPCS	0.386	0.190	0.312	0.122	0.082

Table 5: Overall performance comparison using Precision.

System	$Precision_S$	$Precision_P$
J-ILP (OT)	**0.875**	**0.333**
J-ILP (LT)	**0.875**	0.217
J-ILP (Non-Tran)	0.266	0.002
LexRank	0.350	0.100
LSA	0.375	0.083
KLSUM	0.208	0.042
DSS	0.283	0.053
MRRW	0.325	0.067
LPCS	0.467	0.133

6.5 Parameter Settings in J-ILP

We set the parameters in the proposed J-ILP formulation as follows:

(1) *number of clusters of initially retrieved sentences*: Following [38] we set the number k of latent clusters of each input set in J-ILP formulation as:

$$k = \lceil \sqrt{n} \rceil \tag{22}$$

where n is the number of sentences in each set.

(2) *number of generated pairs for comparison*: In view of the fact that the number of counterparts for each sentence is at most one in the output, we set the number of generated pairs m to be its upper bound $min\{k_A, k_B\}$, where k_A and k_B are the numbers of identified exemplars of the two compared sentence sets.

6.6 Experimental Results

Tab. 4 and Tab. 5 show the performance of summaries generated by all the methods in terms of ROUGE and precision scores, respectively. From the results, we have the following observations. (1) Our proposed J-ILP model with transformation outperforms all the baselines in terms of all metrics. Especially, J-ILP using orthogonal transformation achieves highest ROUGE-1, ROUGE-L, ROUGE-W and ROUGE-SU scores. From Tab. 5, it can be seen that 87.5% sentences and 33.3% sentence pairs generated by J-ILP (OT) are judged as correct by human annotators. These observed results are because the proposed J-ILP formulation takes all the necessary factors (saliency, correspondence and diversity) into consideration. Based on this formulation, the exactly optimal solution can be obtained using the *bound-and-branch* method. (2) Both multi-document summarization methods and comparative summarization methods perform poorly. For multi-document summarization methods, the plausible reason can be that they tend to select very general sentences similar to many other sentences, while such sentences may not have proper counterparts nor may interest annotators. As for comparative summarization methods, they more focus on discovering sentences delivering period-specific information, which prevent them from having proper counterparts as temporally correspondent sentences are information "shared" by different periods.

6.7 Additional Analysis

We further investigate the influence of the parameters used in proposed J-ILP formulation with orthogonal transformation as follows.

6.7.1 Effects of the number of CFTs. We examine now how the performance of orthogonal mapping varies when we change the rate of used CFTs to train the transformation matrix. We test the rate within the range (0, 0.1] and with a step of 0.01. Fig. 1 shows the ROUGE and precision curves of our method, respectively. We can see from the figure that the rate of used CFTs has an effect on the performance of summarization. In this study we set the rate of CFTs as 0.07 based on the observations received from Fig. 1 to obtain the best results.

6.7.2 Effects of trade-off parameter. To clearly show the effect of the trade-off parameter λ in J-ILP for balancing saliency and temporal correspondence, we investigate how the metrics vary per-λ (See Fig. 1). λ is set in the range [0,1] with a step of 0.1. The closer λ is to 0, the less effect the saliency part has. With $\lambda = 0$, J-ILP simply relies on the correspondence of sentences across time. With $\lambda = 1$, the performance merely depends on the representativeness of important sentences without connecting different time periods.

From Fig. 1 we can observe that tuning λ can largely affect system performance. We found 0.4 to perform best, since ROUGE and precision scores w.r.t. all settings reach their maximal values. In general, we can see that λ needs to be fine-tuned to achieve an optimal performance.

6.8 Example Summary

Tab. 6 shows an example of across-time comparative summary generated by using the J-ILP model with orthogonal transformation. The issued query was "Japan". The summary describes some import comparisons between Japan in [1987, 1991] and that of [2002, 2007]. For example, pair (5.1, 5.2) shows Japan's trade connections with different oil exporters at different periods, from Indonesia in [1987, 1991] to Russia in [2002, 2007]. It can be inferred that terms "Indonesia" and "Russia" share similar word vectors after aligning the two compared periods, partly due to their similarity of being main oil exporters at different periods. A similar example

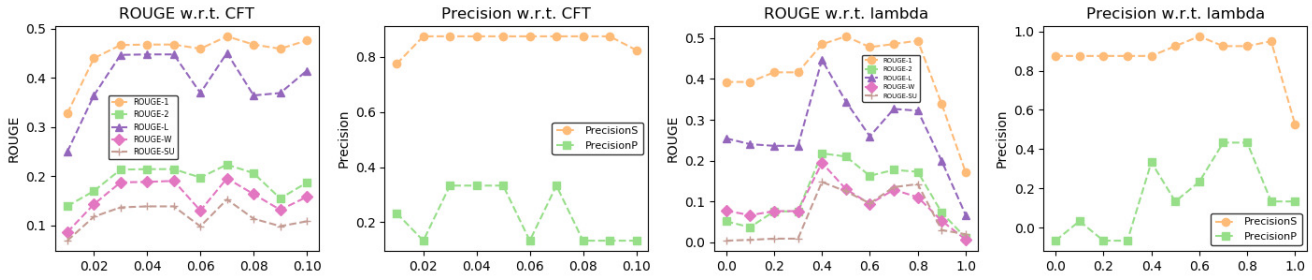

Figure 1: Performance of J-ILP with orthogonal transformation w.r.t. percentage of used CFTs and λ

(10.1, 10.2) reveals that Japan's trade figures with the United States had improved continuously in [1987, 1992], while in [2002, 2007] economic trade seem to be more focused on China. Here, in this particular context "United States" and "China" are considered as temporal analogs in terms of being the main trade partner with Japan. We can also infer other types of comparative knowledge, such as pair (8.1, 8.2) revealing Japanese government's different attitude towards commercial whaling, changing from announcing ending to intending to resume. We can see most of the comparisons are clear and tend to convey temporally correspondent knowledge.

7 CONCLUSIONS AND FUTURE WORK

This work approaches the problem of a special kind of summarization - summarization of past news articles stored in long-term news archives based on user issued queries. The comparative character of our proposed summarization allows finding important comparative aspects in two temporally distant time periods. Users can benefit from such novel kind of access to historical document archives for needs including analyzing trends, determining historical analogies, as well as for educational or entertaining purposes, etc. News archives are constantly growing both due to digitization of born analog articles carried by most of memory institutions (archives, libraries, etc.) as well as due to the ease of generating and exchanging news articles in digital form these days. We believe that novel, effective approaches are needed to make use of large long-term news archives and to benefit from the past experiences of our society.

We first discuss the key challenge of query-sensitive comparative summarization across time and we set up four necessary objectives for realizing effective summarization approaches. We then develop an approach to address this task relying on orthogonal news correspondence measurement and a concise joint integer linear programming framework. Through experiments on the New York Times Annotated Corpus we demonstrate the effectiveness of our approach.

In the future, we plan to test our method using more queries and different time periods. We will also experiment with abstractive summarization methods, which have recently gained more attention in the document summarization community. We are also interested in designing more problem-specific ILP framework with better scalability based on its intrinsic flexibility for the purpose of across-time news summarization and understanding.

Table 6: Example of generated summary by J-ILP with orthogonal transformation (Query: *Japan*). Each row contains a pair of two across-time comparative sentences. The first sentence in each pair is extracted from documents published in [1987, 1991], while the second one is taken from documents published in [2002, 2007].

(1.1) Japan Airlines said it would upgrade its computer reservations and ticketing system to, enable it to link up with American and European systems. (1.2) Japan Airlines said Wednesday that it planned to buy as many as 50 Boeing 7E7 Dreamliner jets.
(2.1) One of every six leading Japanese companies has received extortion threats from organized crime syndicates, a police report released this week said. (2.2) A self-proclaimed corporate raider who struck fear into Japan's insider-run boardrooms by demanding American-style shareholder rights was arrested on Monday on suspicion of insider trading.
(3.1) The Ministry of International Trade and Industry forecast on Friday that Japan's global trade surplus would decline about $18 billion in the 1987 fiscal year. (3.2) Japan's economy contracted sharply in the fourth quarter of 2001, by 1.2 percent, the government reported today.
(4.1) Cecil Fielder, who led the major leagues in 1990 with 51 homers after he played a season in Japan, was among 26 players named yesterday to go to Japan for an eight-game tour next month. (4.2) Japanese soccer officials announced yesterday that its men's national team would not travel ,to the United States for two exhibition games because of the war in Iraq.
(5.1) Pertamina, Indonesia's state-owned oil company, and Japanese buyers have agreed in principle to,a new one-year contract for sales of crude oil. (5.2) Japan is looking to the Russian Far East, ensuring that Sakhalin Island will become a major supplier of oil and gas to Japan within a decade.
(6.1) American and Japanese negotiators met today in the opening round of talks intended to follow through on agreements reached last summer to remove "structural impediments" to trade. (6.2) With President Vicente Fox of Mexico here to sign a free trade pact with Japan, talks broke down Thursday over Japan's dogged defense of its pork and orange juice producers.
(7.1) Japan's Fair Trade Commission said today that its international committee was considering applying anti-monopoly regulations to all foreign companies whose business practices affect Japan. (7.2) The Fair Trade Commission in Japan ruled on Tuesday that the Intel Corporation violated the country's antimonopoly law in the way it sold semiconductors and ordered the company to stop some of its sales practices.
(8.1) Japan recently announced the end to five decades of commercial whaling. (8.2) Japan's latest effort to resume commercial whaling was strongly rebuffed in two votes ,at the biennial meeting of 160 countries adhering to the Convention on Trade in Endangered Species.
(9.1) Foreign car sales in Japan rose 59 percent from last year's levels to a record 9,597 in July, a spokesman for the Japan Automobile Importers Association said. (9.2) Sales at Japan's largest industrial electronics companies rebounded in the October through December quarter on strong demand for optical disk drives, cellphones and the semiconductors used in digital cameras and other hot-selling gadgets.
(10.1) Japan said today that its trade figures with the United States had improved strikingly in the last six months, and it predicted that the trend would continue for the rest of the year. (10.2) Sharply increased trade with China has lifted the Japanese economy out of a lost decade of feeble growth and recurring recession, while cheap imports from China have driven costs down significantly for Japan's long-suffering consumers.

8 ACKNOWLEDGMENTS

This research has been supported by JSPS KAKENHI Grant Numbers #17H01828, #18K19841 and by MIC/SCOPE #171507010, and by Microsoft Research Asia 2018 Collaborative Research Grant.

REFERENCES

[1] Klaus Berberich, Srikanta J Bedathur, Mauro Sozio, and Gerhard Weikum. 2009. Bridging the Terminology Gap in Web Archive Search.. In *WebDB*.

[2] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. 2015. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)* 47, 2 (2015), 15.

[3] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*. ACM, 335–336.

[4] Yllias Chali and Sadid A Hasan. 2012. On the effectiveness of using sentence compression models for query-focused multi-document summarization. *Proceedings of COLING 2012* (2012), 457–474.

[5] Yun-Nung Chen and Florian Metze. 2015. Two-Layer Mutually Reinforced Random Walk for Improved Multi-Party Meeting Summarization. In *Proceedings of SLT*.

[6] Zhicheng Dou, Sha Hu, Kun Chen, Ruihua Song, and Ji-Rong Wen. 2011. Multi-dimensional search result diversification. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 475–484.

[7] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.

[8] Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *EACL*. 462–471.

[9] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315, 5814 (2007), 972–976.

[10] Inc. Gurobi Optimization. 2016. Gurobi Optimizer Reference Manual. (2016). http://www.gurobi.com

[11] Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *NAACL*. 362–370.

[12] William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096* (2016).

[13] Lei He, Wei Li, and Hai Zhuge. 2016. Exploring Differential Topic Models for Comparative Summarization of Scientific Papers.. In *COLING*. 1028–1038.

[14] Xiaojiang Huang, Xiaojun Wan, and Jianguo Xiao. 2011. Comparative news summarization using linear programming. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2*. Association for Computational Linguistics, 648–653.

[15] Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*. IEEE, 229–238.

[16] Amal C Kaluarachchi, Aparna S Varde, Srikanta Bedathur, Gerhard Weikum, Jing Peng, and Anna Feldman. 2010. Incorporating terminology evolution for query translation in text retrieval with association rules. In *CIKM*. ACM, 1789–1792.

[17] Nattiya Kanhabua and Kjetil Nørvåg. 2010. Exploiting time-based synonyms in searching document archives. In *JCDL*. ACM, 79–88.

[18] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515* (2014).

[19] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *WWW*. International World Wide Web Conferences Steering Committee, 625–635.

[20] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*. 957–966.

[21] Yanran Li and Sujian Li. 2014. Query-focused multi-document summarization: Combining a topic model with graph-based semi-supervised learning. In *COLING*. 1197–1207.

[22] Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449, 7163 (2007), 713.

[23] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 71–78.

[24] Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *NAACL HLT*. 250–256.

[25] Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *ACL*. 259–263.

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

[27] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* (2013).

[28] You Ouyang, Wenjie Li, Sujian Li, and Qin Lu. 2011. Applying regression models to query-focused multi-document summarization. *Information Processing & Management* 47, 2 (2011), 227–237.

[29] Mark Pagel, Quentin D Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449, 7163 (2007), 717.

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

[31] Zhaochun Ren and Maarten de Rijke. 2015. Summarizing contrastive themes via hierarchical non-parametric processes. In *SIGIR*. ACM, 93–102.

[32] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[33] Evan Sandhaus. 2008. The New York Times Annotated Corpus Overview. The New York Times Company, Research and Development, 1–22. https://catalog.ldc.upenn.edu/docs/LDC2008T19/new_york_times_annotated_corpus.pdf

[34] Chao Shen and Tao Li. 2011. Learning to rank for query-focused multi-document summarization. In *ICDM*. IEEE, 626–634.

[35] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859* (2017).

[36] Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *ISIM*. 93–100.

[37] Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. 2012. Neer: An unsupervised method for named entity evolution recognition. *COLING* (2012), 2553–2568.

[38] Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *SIGIR*. ACM, 299–306.

[39] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2012. Comparative document summarization via discriminative sentence selection. *TKDD* 6, 3 (2012), 12.

[40] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL HLT*. 1006–1011.

[41] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems* 53, 2 (01 Nov 2017), 297–336. https://doi.org/10.1007/s10115-017-1042-4

[42] Yating Zhang, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka. 2015. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *ACL*, Vol. 1. 645–655.