Int J Digit Libr (2015) 15:169–187
DOI 10.1007/s00799-015-0139-1

Visions and open challenges for a knowledge-based culturomics

Nina Tahmasebi · Lars Borin · Gabriele Capannini · Devdatt Dubhashi · Peter Exner · Markus Forsberg · Gerhard Gossen · Fredrik D. Johansson · Richard Johansson · Mikael Kågebäck · Olof Mogren · Pierre Nugues · Thomas Risse

Received: 30 December 2013 / Revised: 9 January 2015 / Accepted: 14 January 2015 / Published online: 18 February 2015

Abstract The concept of *culturomics* was born out of the availability of massive amounts of textual data and the interest to make sense of cultural and language phenomena over time. Thus far however, culturomics has only made use of, and shown the great potential of, statistical methods. In this paper, we present a vision for a *knowledge-based culturomics* that complements traditional culturomics. We discuss the possibilities and challenges of combining knowledge-based methods with statistical methods and address major challenges that arise due to the nature of the data; diversity of sources, changes in language over time as well as temporal dynamics of information in general. We address all layers needed for knowledge-based culturomics, from natural language processing and relations to summaries and opinions.

Keywords Culturomics · Statistical analysis · Knowledge-based analysis · Temporal text analysis · Digital humanities · eScience · eInfrastructure · Natural language processing

The majority of this work was done while Nina Tahmasebi was employed at Chalmers University of Technology.

N. Tahmasebi (✉) · L. Borin · M. Forsberg · R. Johansson
Språkbanken, University of Gothenburg, Gothenburg, Sweden
e-mail: nina.tahmasebi@gu.se

G. Capannini · D. Dubhashi · F. D. Johansson · M. Kågebäck · O. Mogren
Chalmers University of Technology, Gothenburg, Sweden

P. Exner · P. Nugues
Lund University, Lund, Sweden

G. Gossen · T. Risse
L3S Research Center, Hannover, Germany

1 Introduction

The need to understand and study our culture drives us to read books, explore Web sites and query search engines or encyclopedias for information and answers. Today, with the huge increase of historical documents made available we have a unique opportunity to learn about the past, from the past itself. Using the collections of projects like Project Gutenberg [67] or Google books [25], we can directly access the historical source rather than read modern interpretations. Access is offered online and often minimal effort is necessary for searching and browsing. The increasing availability of digital documents, spread over a long timespan, opens up the possibilities to move beyond the study of individual resources. To study our history, culture and language, we can now computationally analyze the entire set of available documents to reveal information that was previously impossible to access.

The aim of the emerging field *Culturomics*, introduced by Aiden and Michel [3] and Michel et al. [53], is to study human behaviors and cultural trends by analyzing massive amounts of textual data that nowadays are available in digital format. By taking a purely statistical view, Michel et al. [53] unveiled information in Google books that would not have been found without the analysis of this large text corpus. Already by studying word frequencies interesting linguistic and cultural phenomena can be found.

One example of a linguistic phenomenon is the size of the English lexicon. The numbers of unique common words for the years 1900, 1950 and 2000 were compared and the authors found that by year 2000, the size of the lexicon had increased significantly (but see [46]).[1] This means that by the year 2000, more unique words are used in written text

[1] The authors define a common word as one with a frequency greater than one per billion.

than ever before. As an example of a cultural phenomenon the authors studied fame, approximated by the frequency of a person's name. The 50 most famous people born each year between 1800 and 1950 were studied based on several criteria, among others the age of peak celebrity and the half-life of decline. The authors found that, over time, people become famous earlier in their lives and are more famous than ever but are being forgotten faster.

The examples given above showcase the potential of culturomics and could not have been found by studying individual resources or by exclusively using manual analysis. However, despite the large potential of a purely statistical view, culturomics would gain from complementing with deeper, knowledge-based approaches as well as integrate pre-existing knowledge. In this paper, we introduce a knowledge-based culturomics that complements the purely statistical method of classic culturomics with NLP and recent information extraction techniques. We will present our vision and current progress toward a knowledge-based culturomics and discuss open challenges.

1.1 Knowledge-based culturomics

In a broad perspective, culturomics is the study of cultural and linguistic phenomena from large amounts of textual data distributed over a long timespan. Classical culturomics can be used to answer research questions using individual terms—understood as text word types—their frequencies and co-occurrence behaviors.

By "knowledge", we here understand *a priori knowledge relevant to the processing of the material* with these aims in mind. Most fundamentally this is linguistic knowledge about the language(s) present in the material, since the information that we wish to access is conveyed in language, and also general world knowledge and pertinent specific domain knowledge. Using such a priori knowledge allows us to provide additional insights using advanced linking and aggregation of information and relies on a combination of techniques from information extraction and automatic aggregation.

To return to the examples from above, using knowledge-based methods researchers can answer, in addition to the average age of fame, also the typical sentiment toward famous people over time. Have we become more or less positive toward our celebrities? Is there a difference between the start, middle or end of their fame period? What increases a celebrity's fame the most; actions that cause positive or negative reactions? Using *implicit* social graphs, i.e., social graphs extracted using relation extraction and semantic role labeling, we can answer questions about typical social behavior over time. How often do people get married in different parts of the world? Does getting married increase the chances of changing location? Are there differences between different parts of the world? Are people more or less likely to be famous by being married/a child to a famous person?

To answer these type of questions we need a three-layered pipeline: (1) extraction of first-order information; (2) aggregation of information; and (3) establishing connection with the primary resources.

First layer The first layer of processing (not considering the digitization process) is natural language processing (NLP). In this layer, we consider the extracted information to be first-order informational items, among which we include information from lemmatization, morphological, syntactic and lexical semantic analysis, term extraction, named entity recognition and event extraction.

This layer is crucial in our conception of knowledge-based culturomics. Linguistic processing allows for abstraction over text words which in many cases can lower the data requirements considerably, crucial in the case of languages where the amount of available material is smaller, and in the case of morphologically complex languages, and of course doubly crucial when both factors are present.

As an illustrative example of this, in the top diagram in Fig. 1, we replicate one of the examples from Michel et al. [53, p. 180], reproduced here using the *Google Books Ngram Viewer*. This graph shows the appearance of the three surnames *Trotsky*, *Zinovyev*, and *Kamenev* in the Russian portion of the Google Books material, illustrating the rise and fall of these three portal figures of the Bolshevik revolution in the wake of specific historical events.

Now, all examples in [53] are produced on the basis of a single text word form, in this case the base (citation) form of the three surnames. However, Russian nouns (including proper nouns) are inflected in six grammatical case forms (in two numbers), signalling their syntactic role in the sentence. The citation form, the nominative (singular), is mainly used for grammatical subjects, i.e., agents of actions. In the bottom diagram in Fig. 1, we have added the accusative/genitive form of the three surnames,[2] i.e., the form used for direct objects (patients/goals of actions) and possessors. It is obvious from the graph that there are almost as many instances of this form as there are of the nominative form in the material. This shows a clear indication that morphological analysis and lemmatization—using a priori linguistic knowledge—can help us to get more "culturomic leverage" out of our material.[3]

[2] With male surnames, the accusative form and the genitive form are always identical (although otherwise generally distinguished in Russian morphology).

[3] Mann et al. [51] describes an improved version of the Google Books Ngram Viewer where searches can be made for inflected forms of words. However, rather than using a dedicated morphological processor for each language, this project relies on Wiktionary data for collecting inflectional paradigms. Following a general lexicographical tradition, names are generally not listed in Wiktionary. However, names have

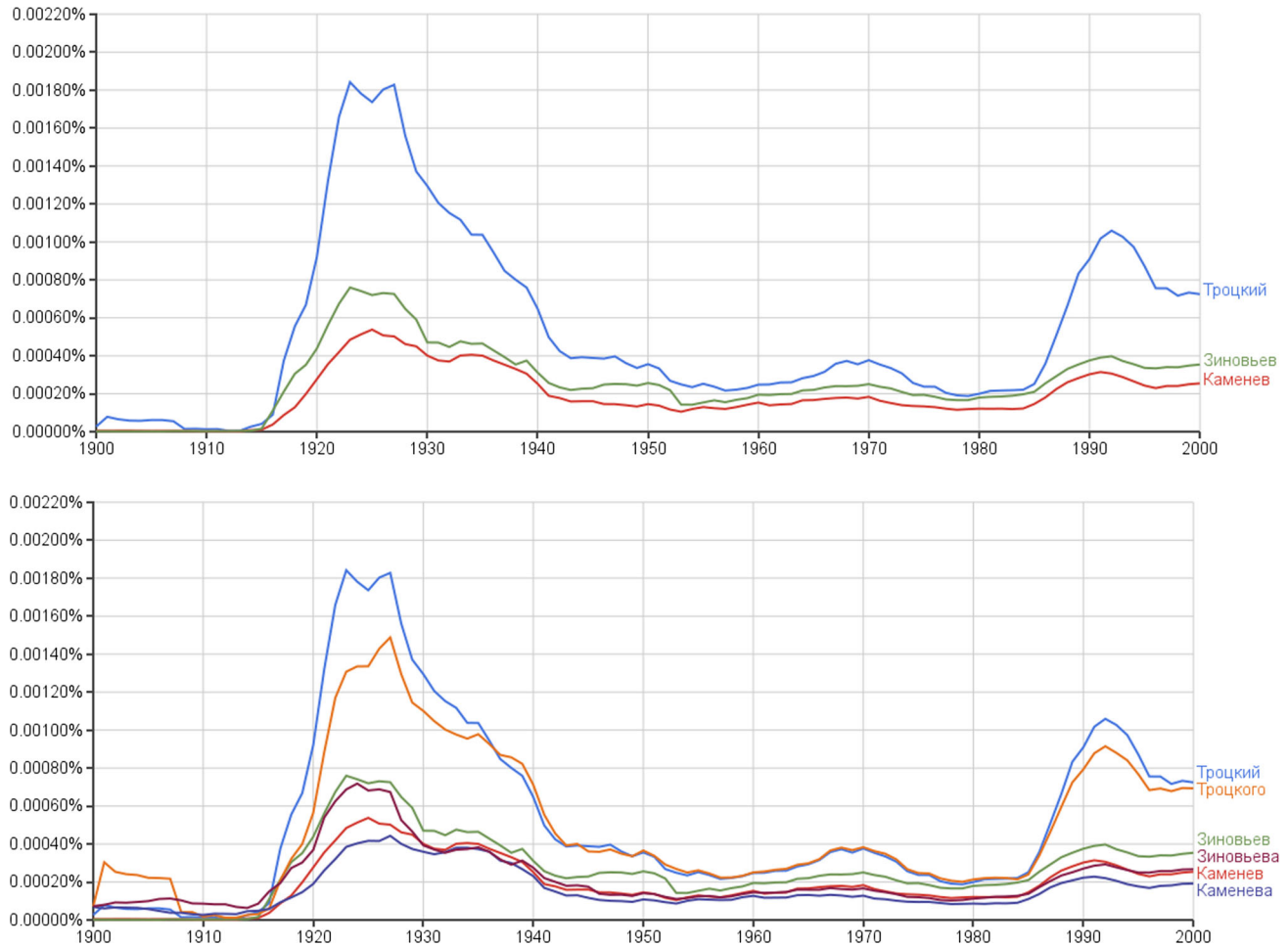

Fig. 1 The Russian surnames *Trotsky*, *Zinovyev*, and *Kamenev* in the Russian portion of the Google Books material, in the nominative case (*top*) and the nominative plus the accusative/genitive case (*bottom*)

A central aspect of knowledge-based culturomics as we construe it is consequently to apply as much linguistic processing to the material as is possible, using the most mature solutions for each language. This is arguably the step which can add most value in relation to the needed effort, and one where mature systems exist for many languages (although the picture tends to look less rosy when it comes to historical language stages or text types such as social media; see below).

In this paper, the technologies on the first layer are presented primarily as background and the focus lies on highlighting the specific challenges for culturomics.

Second layer On the second layer of processing, first-order informational items are aggregated to reveal "latent" information. We consider such information to be second-order information. Entities, events and their semantic roles can be combined to create implicit social graphs where people are related to each other (e.g., *parent-of*, *married-to*, *communicates-with*, *collaborates-with*) or more general entity graphs where, e.g., people can be related to events or locations (e.g., *born-in*).

Topics and opinions play an important role at this layer as they can reveal cultural phenomena. Opinions are extracted, either as general levels of sentiment or as sentiments held by an opinion holder toward a given entity, event or location. Topics can be extracted on general levels to allow exploration or with respect to a person or event. Summaries can be created to help exploration and ease understanding. These summaries can be on the basis of individual documents or documents related to entities, events, topics or opinions. Also language changes are detected on this layer.

Some of the mentioned technologies have been researched in detail during the past decades. However, large chal-

Footnote 3 continued
distinct inflectional paradigms in Russian and many other languages, and consequently, it is still not possible to retrieve all forms of Russian surnames automatically in the Ngram Viewer.

lenges lie ahead when placing these in the context of culturomics. One important challenge lies in including the temporal dimension and monitoring for change and dynamics. Tracking of information and temporal linking over large timespans must, among other things, handle language change. A second challenge lies in handling of historical texts without pre-existing reliable, digital knowledge resources like Wikipedia or WordNet, that sufficiently cover each period of time. As an example, opinion mining relies to a large extent on existing sentiment dictionaries. Such dictionaries exist only for modern content and cannot be generalized to cover the entire timespan without major efforts, and thus, opinion mining for culturomics requires new or adapted methods, as well as the development of new knowledge-rich resources, from scratch and/or by cross-linking from modern resources.

Third layer The aggregated information has the power to reveal patterns and phenomena that could not be detected by studying individual resources. However, the aggregation often brings us to a *distant reading scenario* [59] where we no longer have a direct link to the primary resources. Without access to the individual resources, it is difficult to verify, evaluate or perform detailed analysis of the information. Therefore, the final layer of knowledge-based culturomics is to re-connect extracted information with the primary resources. A large challenge here is to determine for which type of information this linking is at all possible. As an example, the size of the English vocabulary over time cannot be found in any one document in the collection. This information is reflected only in the aggregation and therefore there is no document that directly mentions or reflects this fact. However, relations, events, opinions and word senses might have primary representation. Therefore, this task is concerned with finding the set of most *descriptive* primary resources for a given fact or relation. Where the primary resources are known, the task is to choose the most descriptive resources among all known resources.

1.2 Structure of the paper

The rest of this paper is structured as follows: Sect. 2 covers natural language processing and first-order information such as entities and relations. This section also discusses entity resolution on the basis of extracted information. Sections 3–5 cover the second layer; Sect. 3 covers language change and Sect. 4 covers temporal opinion mining. Temporal semantic summarization is covered in Sect. 5. The third level—finding descriptive resources—is covered in Sect. 6. Related work and discussion are covered in each section individually. In Sect. 7 we present a general discussion on knowledge-based culturomics and finally we conclude the paper and present future work in Sect. 8.

2 First-order information, the NLP layer

The NLP layer is a linguistically motivated information layer added to the texts, which forms a basis for further analyses and allows us to extract first-order information items. Using NLP we can, e.g., identify that *mice* is the plural of the lemma *mouse*, make distinctions between words such as the noun *fly* and the verb *fly*, or determine the syntactic role of the named entity *Michelle Obama*. In this layer, a challenge is entity resolution as well as determining relation between entities.

2.1 Entity resolution

A key component in understanding and analyzing text is to identify and reason about entities mentioned in the text. In general text, the task is made difficult by the fact that entities are not always referred to using their full name, but with a pronoun, a noun phrase or by different names. Furthermore, several entities may have the same name. Merging mentioning of the same entity as well as differentiating between entities with the same lexical reference is an important challenge in culturomics as the entities come from a wide variety of sources and have been created over a long timespan introducing a larger variety.

As an example we can consider measuring the fame of a person over time, an example from Michel et al. [53]. We want to compare the artists *Michael Jackson* and *Madonna*. First, it is important to recognize that *the King of Pop* refers to *Michael Jackson*. However, mentioning of *Michael Jackson* in the 1880s as well as in domains like science, sports and religion is unlikely to refer to the same person and should be disregarded. For *Madonna*, entity resolution is even more important as the highest frequency for *Madonna* in Google books is around year 1910 in reference to the religious figure rather than the artist.

Central to entity resolution is the process of linking pronouns or phrases to entities, commonly called coreference, or entity resolution and is needed when aligning information about entities across documents as well as across time. While humans perform this task seemingly without effort, automating it has proved to be a greater challenge. Nowadays, because gathering massive amounts of text data is done with relative ease, the need for automatic analysis tools, such as extraction and resolution of entities, is increasing.

At its core, the problem of entity resolution is one of linking or grouping different manifestations of an underlying object, e.g., {*Michael Jackson, Michael Joseph Jackson, MJ, the King of Pop*} $\longrightarrow$ *Michael Jackson*. One of the earlier instances is that of record linkage [60] in which multiple database records, of the same object, are merged together. More recently, methods approaching this problem specifically for text have emerged. Focusing first on simple rule-based methods [42,69], the approaches became increasingly

more sophisticated, adopting the tools of machine learning [52] to unsupervised statistical [61] or clustering methods [10,71] and other statistical methods targeting temporal aspects [8,77].

In culturomics, we consider the entity resolution problem (ER) to be that of finding a mapping between references in text to an (unknown) set of underlying entities. Classically, ER is considered within a single document [52]. Leveraging an entire corpus instead involves reconciling entities across documents, increasing the complexity, but improving the analysis. It allows for use of additional features based on document metadata and richer statistical methods for more accurate resolution [9,66].

Rule-based methods are typically deterministic and easy to interpret but require a large set of rules and strong knowledge of the domain. Supervised methods, on the other hand, require labeled data, used for training a model, which may be hard to obtain. Statistical methods build on a set of assumptions on the data, such as an underlying distribution. Often features from rule-based methods are incorporated into statistical methods [30] to enforce linguistic constraints.

Recently, with the advent of global, freely accessible knowledge bases such as DBpedia or YAGO, a new approach to entity resolution has emerged, incorporating world knowledge to aid the process [70]. However, for many historical datasets (or fiction for that matter), such knowledge bases do not contain sufficient information about the entities involved. Another recent trend is to focus on very large scale problems, with 10s or 100s of millions of documents and entities. In such a setting, cheap identification of likely ambiguous identifiers is a helpful tool to avoid unnecessary computations. This problem has been approached by Hermansson et al. [33], using graph kernels and co-occurrence information, to classify identifiers as ambiguous or not.

As a special case of ER we consider temporal resolution, also called named entity evolution recognition (NEER), as the task of linking different names used for the same entity over time, e.g., cities and people and also different underlying concepts, like the *Great War* and *World War I*. For temporal resolution, statistical methods have exploited the *hand over* between different names of the same entity. This method shows great advantage for temporal entity resolution but would benefit from ER as a first step. Methods for temporal resolution of underlying concepts, rather than entities, are not well explored and need to be tackled more in depth for proper culturomics analysis. See further Sect. 3.

2.2 Relation extraction and semantic role labeling

Relation extraction is the task of extracting specific semantic relations between words or phrases, and the entities they refer to. While relation extraction considers mostly binary relations, semantic role labeling (SRL) is an extension to general predicate–argument structures.

Applying semantic role labeling to large corpora enables culturomics to extend the analysis from isolated words or n-grams to predicate–argument structures. Such structures may reveal finer cultural concepts as they exhibit relations between the concepts. The outcome of this analysis feeds extensively into the remaining technologies (e.g., temporal semantic summarization, Sect. 5). Below is a set of frequent triples consisting of a subject, a predicate and an object extracted from Wikipedia using the Athena system [20]:

Males have income
Schools include school
Students attend schools
Couple have children
Teams win championships
Album sell copies

The quality of the found relations affects the remaining methods in knowledge-based culturomics. Therefore, an important challenge of relation extraction is to keep a high quality while being able to scale to large datasets. In particular, for datasets that are diverse and vary over time.

Supervised methods have been used for relation extraction. They usually exhibit high performance in terms of precision/recall. However, they rely on a hand-annotated corpus for training. Since hand labeling is laborious and time consuming, these corpora are often small, making supervised extraction unscalable to Web size relations or large historical corpora. In contrast, distant supervision (DS) presents a novel alternative. It relies on existing relation facts extracted from a knowledge base to detect and label relations in a sentence. DS draws from both supervised and unsupervised methods in having relatively high performance and domain independence, providing canonical relational labels, without losing scalability in terms of documents and relations. DS was first introduced for information extraction (IE) in the biological domain by Craven et al. [17]. Since then, DS has been successfully applied to relation extraction, see [7,13,57,72,91].

In DS, training data are created using heuristic methods by matching entities in text to entities in relations from a knowledge base. For each matching set of entities, a relational tuple is formed by labeling with the corresponding relation from the knowledge base. By extracting features from the matching entities and sentence, a classifier can be trained. Because of the nature of the data, diversity of sources and changes in language, distant supervision seems to be the most appropriate way to go for culturomics.

2.3 Discussions

Since the NLP pipeline provides the first-order information layer for the subsequent analyses, it is crucial that special attention is spent on improving the quality of this information layer—a small error in a previous analysis step tends to multiply in the following ones. An endeavor such as culturomics must hence be informed by the latest developments in NLP research concerning the first-order analysis, in particular, the ones aimed at the target language at hand, instead of just unreflectingly using available off-the-shelf tools.

There are also many linguistic a priori distinctions built into the NLP pipeline that may influence later analyses substantially. These distinctions are concerned with linguistic identity; when are constructs in a language considered the same, and when are they different? This regards spelling variations as well as inflected forms of the same name (see Sect. 1 for an example). Also sense information is important; in addition to spelling variations, how many senses does a word like *color* have,[4] and how can these senses be distinguished computationally? The answers to these kinds of questions will have a significant effect on the subsequent analyses. As an example, Tahmasebi et al. [79] showed that by correcting OCR errors, the number of automatically derived word senses was increased by 24 % over a 201-year timespan and by 61 % in the period 1785–1815 where the amount of OCR errors was the largest.

Relation extraction and semantic role labeling may provide better insights on cultural patterns and their variation over time. FrameNet, an electronic dictionary based on frame semantics [22], the theory behind semantic role labeling, explicitly aims at building a repository of shared mental concepts. However, these semantic analyses are still less accurate than tagging or syntactic parsing as they come at the end of the NLP processing pipeline. In addition, they require predicate–argument dictionaries and annotated corpora that are, at writing time, only available for few languages: English, Chinese, German, Spanish, or Japanese and have poor coverage of historical variations.

The context of culturomics presents several new challenges to existing methods for entity resolution. First, the corpora involved have long timespans which make methods based on linguistic rules hard to employ, without allowing for evolving rules. While this problem has been approached in part [85], it is yet to be applied to historical data. Second, the large scale of many historical corpora makes supervised methods hard to use because of the amounts of annotated data needed for accurate classification. Also, it is not clear how the time dimension affects the results of existing methods. Unsupervised methods seem well suited, aiming to discover the underlying rules of the data, rather than state them. To the best of our knowledge, however, no such model exists, targeted specifically to historical data.

[4] The Princeton Wordnet posits 14 senses for the word *color*, of which seven are nouns, six are verbs, and one adjective.

3 Language change and variation

Change and variation are inevitable features of our language. With new inventions, changes in culture or major events, our language changes. Mostly, we are aware of all contemporary changes and can easily adapt our language use. However, over time, linguistic expressions and constructions fall out of use and are no longer a part of our collective memory. For most everyday tasks, this does not cause problems. However, in culturomics, when looking to the past, trying to make sense of cultural and language phenomena over time, recovering past language change is highly important.

There are different types of language change that we consider. The classification depends on how each type affects finding (i.e., information retrieval) and understanding of a given document.

The first type of change is *spelling variation* where words are spelled differently over time. To find the true frequency of a word, all different spellings must be found and the frequencies merged. For example, *infynyt*, *infinit*, *infinyte*, *infynit*, *infineit* are all historical spelling variants used at different times for the word now spelled *infinite*. To follow this controversial concept over time, frequencies and contexts from all spellings must be taken into account.

The second class of language change is *term to term evolution* or more generally *word to word evolution*. Different words are used to refer to the same concepts, people, places, etc. over time. To find the mentioning of such a concept, all different temporal references must be found. Because references do not need any lexical or phonetic overlap, this class is separate from spelling variation. Examples include *The Great War* and *World War I* that refer to the same war (i.e., underlying concept) or *St. Petersburg*, *Petrograd* and *Leningrad* that all refer to the same city (i.e., same entity).

The third class of change is *word sense evolution* or *semantic change*. Words change their meanings over time by adding, changing or removing senses. This means that even if a given word exists across the entire timespan, there is no guarantee that the word was always used to mean the same thing. As an example, assume that we are looking for *awesome leaders* over time. They are likely to appear today as well as several centuries ago; however, their interpretation over time has changed.

For normal users, not being aware of changes can limit their possibilities to find relevant information as well as interpret that information. As an example, not knowing of the name *Petrograd* or *Leningrad* will limit the amount of information that can be found on the history of the city. In the

context of knowledge-based culturomics, language change causes an additional set of problems. As an example, alignment of topics over time requires consolidating words and entities that represent the same concepts over time.

The classes of language change mentioned above fall into two general categories. Spelling, word to word as well as named entity evolution all fall into one category where the same concept or entity is represented using several different words over time. In contrast, word sense evolution falls into a different category where the word is stable over time but the senses change.

While the latter category is important for understanding of text, the former category of change is most relevant for knowledge-based culturomics and has a high impact on other technologies mentioned in this paper.

So far, spelling variation is the class that has received most attention by researchers and is the class that is best understood from a computational point of view. The work has focused on developing automatic/semi-automatic methods using rules as well as machine learning for creating dictionaries and mappings of outdated spelling. These resources are then used for search in historical archives to avoid missing out on important information [1,2,19,26,32].

Named entity evolution has been targeted using statistical methods as well as rule-based methods. Berberich et al. [8] proposes reformulating a query into terms prevalent in the past and measure the degree of relatedness between two terms when used at different times by comparing the contexts as captured by co-occurrence statistics. Kaluarachchi et al. [37] proposes to discover semantically identical concepts (=named entities) used at different time periods using association rule mining to associate distinct entities to events. Two entities are considered semantically related if their associated event is the same and the event occurs multiple times in a document archive. Tahmasebi et al. [77] makes use of special properties, namely *change periods* and *hand overs*, to find named entity evolution. If an entity is of general interest, then the name change will be of general interest as well during the period of change and the two names will be first-order co-occurrences in the text. E.g., *Sture Bergwall is better known to most people in Sweden as Thomas Quick*, ….[5]

For other classes of language change, these properties do not necessarily hold. Words used to refer to the same concepts, i.e., word to word evolution (see Fig. 2), are not likely to be first-order co-occurrences in a text and therefore more difficult to detect. E.g., in the past the word *fine* has been used to refer to the same concept as the modern word *foolish*. However, because there was no hand over as with the *Sture Bergwall* example above, the same methodology cannot be used to connect *fine* and *foolish*. Instead, to find general

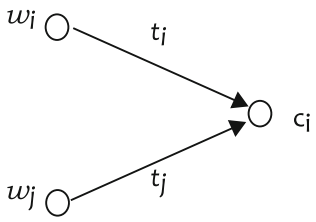

Fig. 2 Words w_i and w_j are considered temporal synonyms or word to word evolutions because they represent the same concept c_i at different points in time

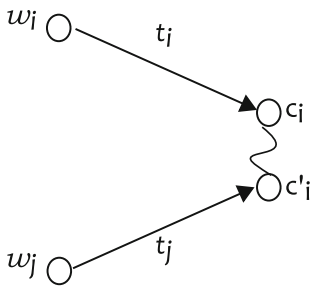

Fig. 3 When the gap between time points t_i and t_j is large, also the words constituting the concept are likely to change and hence we must be able to connect two concepts over time to find word to word evolution

word to word evolution, word senses must be used. A word is first mapped to a concept representing one of its word senses at one point in time. If one or several other words point to the same concept largely at the same period in time, then the words can be considered synonyms. If the time periods do not overlap, or only overlap in a shorter period, the words can be considered as *temporal synonyms* or word to word evolutions.

A key aspect to this methodology is the approximation of word senses. Because we cannot fully rely on existing contemporary dictionaries or other lexical resources, we must find these word senses automatically. Tahmasebi et al. [78] used word sense discrimination to approximate word senses. The curvature clustering algorithm was used and the output verified to correspond to word senses also for English text from the nineteenth century. Though the quality of the found word senses was high, comparably few word senses could be found. In other works, context-based approaches have been used [75] as well as methods making use of probabilistic topic models [40,88]. With the exception of Mitra et al. [58], Tahmasebi [78] and Wijaya and Yeniterzi [88], senses have not been followed over time. This is a key challenge, in particular for culturomics with large timespans, as very few senses remain exactly the same over longer periods of time, see Fig. 3. As words are exchanged in the vocabulary, also words that constitute the senses are exchanged. Therefore, to find word to word evolution, both word senses and the tracking of word senses must be handled automatically.

[5] http://www.gq.com/news-politics/newsmakers/201308/thomas-quick-serial-killer-august-2013.

3.1 Discussion

In culturomics, language change is a great obstacle for most other tasks like aligning and tracking opinions and topics or creating summaries.

Language change is interesting on its own; are there periods in which we change our language more than others? Is language change correlated to major events or cultural changes, e.g., introduction and wide spread use of Web log brought with it a whole set of new terms like *blog*, *blogging*, *blogger*. Are there other such events? Which type of change is more likely to be stable and which type of change is ephemeral and how have these classes changed over time? Did the introduction of user-generated content increase the amount of ephemeral changes?

The technical challenges lie in the lack of datasets and methods for automatic testing and evaluation over long timespans. Typically, tasks like word sense discrimination or named entity evolution have been developed and evaluated on short, modern timespans. While certain tasks are simpler to evaluate (e.g., named entity evolution) tasks like word sense discrimination or evolution are significantly harder, in particular when long timespans are taken into account and manual evaluation becomes infeasible.

4 Temporal opinion mining

Opinions and sentiments are an important part of our culture and therefore a central tool for analyzing cultural phenomena over time. The goal of opinion mining is to detect and classify expressions of approval or disapproval about a topic or entity from a text. The detected opinions can be used on their own to find documents expressing a given point of view, or aggregated to find opinions toward a given topic or event. Aggregation takes place both on the granularity of the topic and on the opinions contributed by different authors.

Research questions like: *are people generally more happy* can be answered by analyzing expressed opinions rather than counting the frequency of the word *happy*.

Currently, aggregation of opinions is typically only done for a single topic on a rather homogeneous dataset in a relatively short time period. However, we can gain more insights if we look at opinions over a longer period of time. In particular, recent years have seen much interest in applying opinion extraction techniques to texts from social media such as Twitter, which allows for a fast, realtime tracking of political sentiment developments [63]. Analyzing the temporal dynamics of these opinions, e.g., toward election candidates and their positions, can be useful for understanding the trends of political attitudes over time [18]. Furthermore, we can also extend this tracking to other cultural and social issues.

When we deal with historical collections, we will first have to analyze which documents contain opinions (e.g., personal letters and editorials) and how those opinions are expressed. Furthermore, the timescales in historical collections are larger and the number of documents is smaller. Whereas we can find opinion events on social media that last only some days or weeks and still contain opinions from thousands of users, the opinion events available in historical collections last for years or even decades while only having a small number of sources. This requires a more careful aggregation strategy that should also take into account external knowledge such as relations between the authors' contributing opinions.

Finding opinion expressions in text is a hard problem. On the one hand, we have texts from social media where the messages are typically very short and often written in a very informal style, which makes NLP analysis harder than for traditional texts [28]. On the other hand, we have media such as newspapers that often contain opposing opinions from different parties in the same text, sometimes adding an author's viewpoint as well. The problem then is to correctly attribute each opinion to the correct speaker, similar to the problems of entity resolution discussed in Sect. 2.1. Common additional problems are domain-specific vocabulary and slang, as well as the use of irony and sarcasm, which make it harder to find the intended meaning.

4.1 Related work

Opinion extraction (or "opinion mining", "sentiment analysis", or "subjectivity analysis") is a wide and diverse field of research [50,65]. Automatic retrieval of opinionated pieces of text may be carried out on a number of different levels of granularity. On the coarsest level, *documents* are categorized as opinionated or factual [90]. This may for instance be used to distinguish editorials from news [93]. Classification of *sentences* is also widely studied; these classifiers have been based on linguistic cues [87] or bag-of-word representations [64]. While most work use supervised machine learning, there are also unsupervised approaches [41,48].

In contrast to the early work, recent years have seen a shift toward more detailed and fine-grained problem formulations where the task is not only to find the text expressing the opinion, but also analyzing it: *who* holds the opinion (the holder) and toward *what* is it directed (the target); is it positive or negative (polarity); what is its intensity [16,35,39,89]. The increasing complexity of representation leads us from retrieval and categorization deep into natural language processing territory; the methods employed here have been inspired by information extraction and semantic role labeling, combinatorial optimization and structured machine learning. For such tasks, deeper representations of linguistic structure have seen more use than in the coarse-

grained case [27,35,39,73,76]. An extensive survey of opinion mining is provided by Pang and Lee [65] and more current developments are summarized by Tsytsarau and Palpanas [82].

4.2 Opinion mining for culturomics

The source documents for culturomics are very diverse. They span a large time period, range of registers and are written by authors of widely different backgrounds for varying purposes. As a result, the language as well as the kind of opinions expressed vary. We will describe both aspects as well as proposed solutions in the following.

The language in the documents is different in the diachronic as well as synchronic axis. Diachronically, we have the problem of language change (see Sect. 3), where the connotations of terms change over time. For example, the word *gay* is used to express a positive sentiment (*cheerful*), but now expresses a neutral or in some contexts even a negative sentiment. Similarly, *euphemism treadmill* describes the phenomenon that terms intended as neutral substitutes for negative or "taboo" terms soon become negative themselves, a process that is often iterated as for example in the series *Negro*, *black*, and *African-American*. Synchronically, the language used in texts confers different opinions based on the speaker, topic and type of document. Continuing the example given above, *gay* is used as an approximately neutral synonym for *homosexual* in most Western newspapers, but as a pejorative in many youth cultures. As another example, *primitive* is typically used negatively, but in the context of art it is used neutral and descriptive.

Many opinion mining methods incorporate polarity dictionaries, i.e., lists of words with an a priori known opinion value. As we have however seen, these dictionaries will have a large error when applied to diverse collections. Therefore, it is necessary to automatically adapt the dictionaries to the processed texts.

Prior work has shown that it is possible to generate and extend polarity dictionaries in an unsupervised manner using grammatical [31] or co-occurrence relations [83] between words. By applying these methods on Web data, we can also infer the polarity for slang and common misspellings [84], which improves the quality of opinion mining on, e.g., social media data. The mentioned algorithms build word graphs, where the edges indicate that there is a correlation or anti-correlation of the corresponding node words, e.g., because the terms occur in a conjunction "good *and* cheap" or disjunction "good *but* expensive". Each node is assigned a polarity by propagating labels starting from a seed set of terms with known polarity, taking into account the edges between terms.

To improve the quality of the generated dictionaries, we can incorporate additional metadata such as the document type and the temporal context of the source documents into the graph creation context. For example, we can split nodes for terms that had a change in meaning and thus learn the polarity of the word in different time periods. In this way, we derive a polarity dictionary that contains information about the temporal and contextual validity of its entries.

Another way to improve the accuracy of opinion mining in heterogeneous collections is to use intra-document relations between contained entities and terms (see Sect. 2.2). Current methods only consider the local context of the opinion expression for the classification. By building a global opinion model for a document with the speakers and entities contained in it, we can classify the opinions in that document from this global perspective and correct for local misclassifications from incorrect models and dictionaries. This makes the opinion classifier more robust against heterogeneity in time and type of the input documents.

4.3 Opinion aggregation

The opinions expressed in a collection express the values of the members of the cultures that created the documents. By aggregating the opinions about specific topics we can gain insights into these values and thus understand and characterize the culture. E.g., what can be said about the general levels of satisfaction (as expressed in written documents) in different countries post-World War II? Were there different change rates for countries/regions? Was there a correlation with winning vs. loosing sides?

Opinion aggregation is a hard problem, as we not only have to find all relevant viewpoints, similar to what we need to do in relation extraction, but also need to judge how representative these opinions are. Here, we have the problem that documents are primarily created by people having strong opinions on a topic and are therefore biased, whereas other people may not find the topic relevant enough to express their opinions in written form [14]. Historically, common folk were not even offered the possibility to express themselves in published media, leaving their voices less heard today. Traditionally, the venue for publication has been used as an indicator of the relevance of an opinion, so that the opinion expressed in an article in the New York Times was deemed more important than one published in a local newspaper. However, the Internet and especially social media can help us gather opinions from a larger sample of people instead of only this obviously biased subset, though caution must be taken when gathering data to avoid introducing new bias by, e.g., oversampling from western, internet using teenagers.

Current research is on finding influential users and detecting the spread of information in online social networks either through the network structure (e.g., [5,15]) or through the linguistic influence [43]. These methods can give us a way to approximate the influence of an author's opinion toward the overall culture.

4.4 Opinion dynamics

When our collection spans a longer period of time, we can in addition to aggregating opinions also analyze the dynamics of the aggregate sentiment and see evidence for changes in the value system. On the one hand, we can analyze changes in the polarity, for example, in the confidence in political or social institutions. On the other hand, we can also observe changes in the overall importance to the culture, for example of religion, by looking at the changes in the total number of opinions expressed.

Opinion changes belong to two distinct categories: first are changes that occur in response to an extra-ordinary event. As an example, winning an important contest will prompt more positive opinions about an athlete. The second category of change contains slow but continuous shifts in the aggregated opinion. Continuing the previous example, a continuous series of losses would cause the athlete to slowly lose favor. It is possible for an entity or topic to experience both types of changes at different times. For example, the European sentiment toward the Euro currency had been growing more positive continuously until the financial crisis, which then caused flare-ups of anti-Euro sentiment in multiple countries.

Opinion changes of the first kind can be detected using the associated *opinion change event*. Existing approaches to opinion change detection therefore rely on existing event detection algorithms to detect such events, either indirectly by finding events and analyzing the opinions in their temporal context [81] or directly by detecting changes in the aggregated opinions [6,62].

Slow changes are harder to detect, as they are typically more implicit. For example, a change in the attitudes toward foreigners will only be partially observable through opinion expressions about "foreigners" per se, but rather through changes in aggregate of opinion expressions about individuals seen as members of that group. Therefore, it is necessary to aggregate opinions about groups of related entities and analyze the dynamics of the aggregated sentiment.

4.5 Discussion

Opinion mining can help to answer many culturomics research questions by providing insight into the opinions and values expressed in a document collection. The unique challenges of culturomics, such as the diversity of document sources and topics, as well as the longer time periods, have however not been tackled in previous work. As the techniques for opinion mining mature, these challenges will need to be addressed, especially as they also increase the robustness of opinion mining for more general applications.

A particular area where knowledge-based culturomics can help to drive further research is the detection of slow changes in cultural opinions and values. These manifest themselves through opinions expressed toward groups of related entities from different time periods that form the topic of the opinion change topic. Tools like automatic relation extraction (see Sect. 2.2) and Named Entity Evolution Recognition (see Sect. 3) can help us find the relevant entities, so that we can find all relevant opinions and analyze their dynamics. Furthermore, key resource finding (see Sect. 6) can help us corroborate the detected changes with relevant source documents.

5 Temporal semantic summarization

Automatic summarization is the process of producing a limited number of sentences that outline a set of documents and constitute a summary (Fig. 4). The summary should cover the most important topics and avoid redundancy. Automatic summarization is helpful for preventing information overload and can allow people to quickly digest a large set of documents by extracting the most useful information from them. In the context of culturomics, the goal is to make a digest of a chronological course of events, spanning some period of time, gathered from many different sources. Since the input will come from several different sources, this is called multi-document summarization, and summarizing one single document will be considered as a special case.

Because of the long timespans in culturomics, there is a need to go beyond traditional summarization and extend the summaries to temporal summaries, taking into account changes over time. In addition, a good summary needs awareness of important entities, relations and events to fully qualify into knowledge-based culturomics. Therefore, we envision *temporal semantic summarization*, a novel approach making use of open-domain information extraction to extract entities, relations and events from the text to create better and more complete summaries. In the event extraction phase, a temporal analysis will take place, making the summaries temporally aware.

5.1 Related work

Previous work can roughly be divided into three different categories: extractive methods, abstractive methods, and orthogonally, methods that use information extraction.

In *extractive summarization* every sentence in the input documents is considered as a candidate. The task is to choose the most representative sentences, according to some metrics, to include in the summary. A common approach is to represent each sentence as a weighted vector of TF*IDF terms, and compute the cosine similarity between all pairs. Some extractive approaches start by deriving representations of the topics, whereafter the sentences are scored based on *impor-*

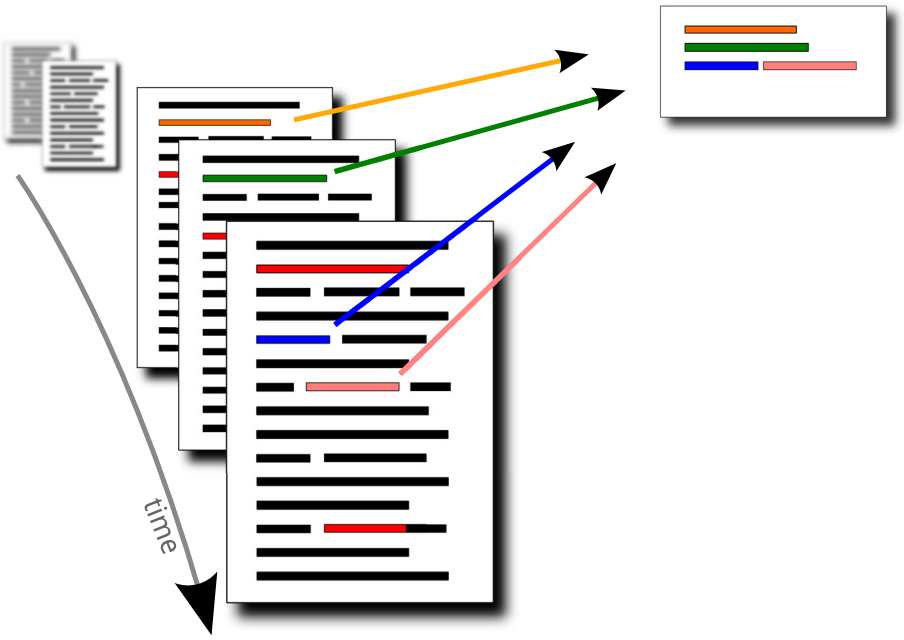

Fig. 4 Illustration of (Extractive) Automatic Summarization

tance. In other methods, the problem is seen as a trade-off between *diversity* and *similarity*. See [47,54,68] for more details.

In *abstractive summarization*, summaries are created by generating natural language sentences. All documents are parsed to build up knowledge in some representation and this knowledge is used to generate new sentences that summarize the documents. Ganesan et al. [23], Leskovec et al. [44], Rusu et al. [74] extracted subject, verb, object triplets from the documents to create a graph based on the semantic relations found. They use different techniques to extract important parts of this graph and generate the summary based on each part. While *abstractive summarization* has great potential, the inherent difficulties with changing language and large amount of noise in the collections used for culturomics leave abstractive methods for future work.

Recently, there has been some focus on using *information extraction to improve summarization*. Filatova and Hatzivassiloglou [21] presented a method that boosts event extraction coverage by relaxing the requirements for an event. Any pair of entities that appear in the same sentence and in combination with a connector word is considered an event. The connector word can be a verb or an action noun, as specified by WordNet. The sentences are scored using a greedy optimization algorithm and the results improve over an extractive baseline summarizer that did not account for redundancy.

Hachey [29] presented *General Relation Extraction*. To classify two entities as participating in a relation, they require the entities to be separated by no more than two words or by only one edge in a dependency parse tree. Connector words are derived from a model of relation types based on latent Dirichlet allocation, avoiding dependency on domain-specific resources like WordNet. The results were similar to those of Filatova and Hatzivassiloglou.

Ji et al. [34] used information extraction to perform relevance estimation and redundancy removal. This was done by combining the scores computed based on IE with scores based on coverage of bi-grams from the extractive summarizer used [24]. Combining these in the right way helped to create summaries that improved over the baselines.

All the above methods show an increased performance over an *extractive summarization* baseline. However, none perform any analysis on temporal information. Temporal summarization has been targeted in the past with a recent upswing [4,11,92]. Allan et al. [4] choose one sentence from each event within a news topic. Yan et al. [92] creates individual but correlated summaries on each date from a time-stamped collection of Web documents. Binh Tran [11] first ranks and chooses the top time points and then chooses the top sentences for each of the chosen time point. All works make use of news articles where typically one document describes one event and timestamps are exact and narrow. Automatic event detection or relation extraction was not considered. To achieve high-quality temporal summaries on historical texts with long timespans, temporal summarization techniques must be combined with information extraction.

5.2 Vision for temporal semantic summarization

For knowledge-based culturomics we envision temporal semantic summaries that build on the previous efforts [21, 29,34] in utilizing IE to create better extractive summaries. In the event extraction step, emphasis on extracting temporal information is needed, allowing for the final summary to be

coherent and chronologically structured. This is in particular important when documents are not properly timestamped, e.g., user-generated content from the Web or old books where timestamps are not fine-grained, alternatively events discussed can refer to historical events and do not correspond to time of publication.

By generating an entity–relation graph, extracted entities can be scored on importance, where importance is temporally variant. This allows to give higher scores to sentences that mention important entities given a time period. It is also possible to compare entity–relation graphs corresponding to different time periods to capture important changes and require these changes to be present in the summaries.

As an example, consider the life of *Marie Antoinette*. Born *Maria Antonia* in Austria as the daughter of the Holy Roman Emperor Francis I and Empress Maria Theresa. As a teenager she moved to France to marry Louis-Auguste and become the Dauphin of France and eventually the Queen of France. Modeling the entity–relation graph would show us significant changes over time as titles, place of residence and strongest connection (i.e., from parents to husband) change over time; also the general opinion toward her changed from being popular to being despised and in modern times again changed in a positive way. These general changes are important for a summary of her life and can be captured by a semantic temporal summarization approach.

Consider newspaper articles describing the world around us. They describe everything from people, events and conflicts to terrorist attacks, summits of world leaders, football stars scoring against some opposing team and companies changing leadership. Using information extraction (IE) to find important entities, relations and events in the input text gives many benefits in the summarization process. To mention a few:

1. Entity and relation extraction helps decide which sentences contain important information. Entity resolution and temporal resolution help to discriminate and consolidate mentioning of same and different entities. E.g., connecting *Maria Antonia* with *Marie Antoinette*.
2. Event extraction gives a natural way of filtering redundant mentions of an event (detecting, e.g., that a *bombing*, is the same as a *terrorist attack*).
3. Information extraction will ideally provide temporal information about events, helping us to order things in the output and disambiguate events.
4. Major changes in entity–relations or changes in events and their descriptions as found by comparing second-order information will provide a measure of importance and improve quality of the summaries.

The first step is extracting the necessary information. This includes building up an entity–relation graph of different kinds of entities (see below) and finding the events. The events can directly or indirectly (and to differing extent) relate to the entities that have been extracted. We will decide on entities in the entity–relation graph that are important, based on how they take part in events and based on the structure of the entity–relation graph.

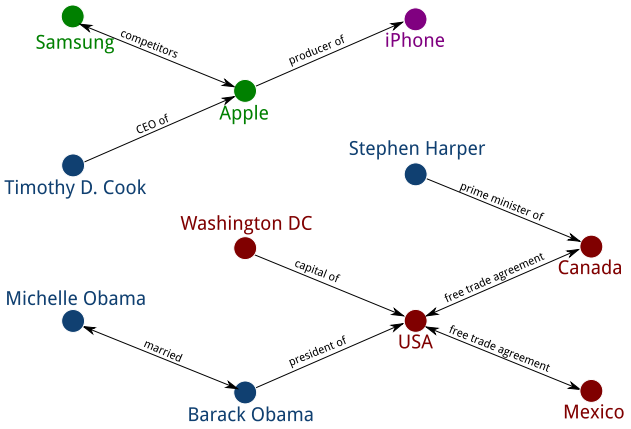

Fig. 5 Example of an entity–relation graph

Once the information is extracted, sentences can be scored based on what information they include. At this point in the process, the objective is similar to that of traditional extractive document summarization. The goal is to include sentences that are representative and relevant. One possibility is to use submodular optimization [47] to select the sentences based on the scores that they received in previous steps.

The process will make use of the following parts.

1. *Anaphora resolution* An anaphora is an expression that refers to some other expression, the most typical example being pronouns referring to some entity within the same sentence, or to other closely located sentences. Anaphora resolution is the process of resolving these expressions.
2. *Entity extraction* Entity extraction is the process of extracting entities from the text. Entities can be people, corporations and geo-political entities.
3. *Entity–relation graph extraction* A graph of relations will be extracted from the input documents (see Fig 5 and Sect. 2). This graph will contain relations between various kinds of entities, such as people, corporations and geo-political entities. Hence, a social graph could be seen as a subgraph of the entity–relation graph, containing only people and the relations between them. The system will also utilize properties from an entity–relation graph extracted from the entire corpus. From this, the aim is to compute a measure of global importance of different entities that are mentioned in the input text. This can be done by applying variants of the page rank algorithm to the network.

4. *Event extraction with temporal analysis* Events can happen once or repeatedly. They can also be hierarchical, with several smaller subevents spread over time, comprising one bigger event.

An entity in the entity–relation graph can be touched by a series of different events. This should be reflected in a summary, hence the temporal aspect of the event extraction is central.

5.3 Discussion

Temporal semantic summarization can help users deal with information overload, and follow developments within topics of interest over lengthier periods of time.

As the method of scoring information in the documents will rely on the success of extracting information, a great challenge lies in combining all these techniques in the best possible way. The aim is to develop a system that works on big amounts of input text, an aspect that can be seen both as a challenge and a source of strength. Algorithms need to be able to cope with the volume, variety and velocity of large amounts of data as well as be robust against noise (e.g., OCR errors or spelling variations in text).

An important aspect of a good summary in a culturomics context is *completeness* as well as *order*. A summary should contain at least the major events (e.g., marriage, mergers, battles) and the strongest relations (e.g., people, places and titles), all sorted in time. For temporal ordering, an important aspect becomes to sort information from documents written about an event during the event compared to information found in documents written after a significant time has passed. To sort the latter type of information, publication time cannot be used, instead time references must be extracted from the text itself, introducing an additional difficulty.

Finally, a proper summary should contain not only facts but also different viewpoints. Relying on the output of the opinion mining step, methods should be found to bring the opinions into the summary, also these in a temporal order. E.g., what were the different viewpoints of the Vietnam war during the war compared to the viewpoints after the war, and in particular, what are the differences.

An interesting addition to both summaries (and opinions) is the differences between automatically created summaries using information available during an event compared to information available post-event. What has been learnt afterwards and what has been forgotten? Which events or relations were deemed important in hindsight that were not considered important during the course of an event and vice versa?

6 Key resource finding

The primary focus within culturomics is the extraction of knowledge. Unfortunately, this is typically accomplished at the expense of losing the connection to the primary resources from where the knowledge arose. In some cases, this is unavoidable, e.g., corpus word frequencies can never be linked to any one primary resource but live in the combined space of all resources. Still, there are other instances where this may be remedied, e.g., relations, events, word senses, and summaries. In these cases *key resource finding*, in short KRF, which is the third processing layer in knowledge-based culturomics, can contribute credibility and traceability, with a secure basis in facts. In cases where all primary resources from where the knowledge arose are known, KRF is the task of choosing the most representative resources to avoid information overflow.

6.1 KRF and relations

By extracting entities from the text and inferring their relation, an entity–relation graph may be constructed, see Fig. 5 and Sect. 2.2. Given such a set of relations, the task of key resource finding is to find the minimal set of documents that best provide evidence of the correctness (or incorrectness) of the relations.

A simple example is finding a document describing some unspecified relation between two known entities, found for example using co-occurrence statistics or relation extraction. In the case of relation extraction, the relation is known and key resource finding is a straightforward search for documents containing the two entities exhibiting some relation. In the case where the two entities have many relations and where the user is only interested in one, the search is further constrained by specifying the relation. In a case where the relation is unknown (e.g., entities are often found in close proximity within text), key resource finding is the task of grouping documents according to different relations and choosing representative documents from each group.

More generally, a user may be interested in documents describing a complete entity–relation graph. In this case, the search turns in to the optimization problem of finding the minimum set of documents that cover the complete graph. This optimization may be further constrained by the number and length of documents that may be included. Also, the graph may be made more general and include, e.g., events.

6.2 KRF for entity disambiguation

If several entities share the same name, this will lead to ambiguity with regard to the identity of the referenced entity. Entity disambiguation aims to find and resolve this ambiguity. Using methodology described in Sect. 2.1 it is possible to automatically identify entities that are likely to be ambiguous.

Using a graph model like that presented by Hermansson et al. [33], the local graph structure surrounding the ambiguous

entity can be used and cliques can be identified. Then, key resource finding is the task of finding representative documents for each clique, which may be used both to resolve the ambiguity and to describe the disambiguated entities.

6.3 KRF for word senses

Clusters of words are used to approximate word senses, see Sect. 3 for a detailed discussion. These word senses can then be used to track meanings of words over time as well as find words that can be considered temporal synonyms. Concretely, a word sense for a given word is a bag-of-word representation of its context.[6] For example, the concept *music* can be described by words like *singing*, *guitar* and *Spotify*. However, automatically derived clusters lack labels on the word senses leaving it up to the user to interpret the meaning of the cluster from its words. This activity is both tedious and to some extent biased and would greatly benefit from being automated.

For labeling or specification of word senses, key resource finding is the task of finding the minimum set of documents (or sentences) that cover the largest amount of words among the describing words, preferably with linguistic relations between the words. Subsequently, documents or sentences can be ranked on informativeness. Presenting a few good sentences to the user, where the word sense is utilized. E.g., *Music is an art form whose medium is sound and silence.*[7]

An alternative method is to use the continuous space word representations presented in Mikolov et al. [55], infer categorical information regarding the words in the cluster, and use the most common category as label.

6.4 KRF for summarization

The objective of automatic summarization is to construct a set length summary that covers a maximal subset of the information conveyed in a set of documents. This can be done either by picking descriptive sentences (extractive summarization) or by constructing an abstract representation of the information contained in the text and subsequently generating a summary using natural language generation (abstractive summarization). More on summarization in Sect. 5.

Key resource finding constitutes the task of finding the smallest set of documents that cover the information in the derived summary to avoid redundancy and provide users with more details around the important elements covered in the summary. This can be accomplished using techniques similar to what is described in Sect. 6.1, with the summary as query.

[6] A bag of words is an unordered set containing all words represented in the original data.

[7] http://en.wikipedia.org/wiki/Music.

6.5 Related work

Key resource finding is an NLP task that falls under the category of information retrieval (IR). Traditional IR systems score the relevance of documents using keyword-based matching with respect to a given free text query. This approach has the advantage of being tractable for large-scale systems, but does not use higher level information in the text. Using semantic role labeling (SRL), it is possible to construct entity–relation graphs, that subsequently may be used to improve the precision of the search results.

Lin et al. [49] introduced an event-based information retrieval approach that finds relevant documents using entities and events extracted from the corpus. The query is created by the user in a structured form, but is limited to a single predicate. This approach could be generalized to the problem of finding a document describing a well-defined relation, as defined in Sect. 6.1.

A more general approach is employed by Kawahara et al. [38], where a predicate–argument graph is used to represent both the query and the documents. Document selection is done using binary operators (i.e., AND OR) on the predicate–argument structures as defined in the query. However, the method leaves SLR-based ranking as future work and does not provide a way to determine how many (or which) documents that are needed to cover the complete query, where the latter is necessary to answer the general question posed in Sect. 6.1.

6.6 Discussion

An important aspect of key resource finding is to step away from exact matching of words and sentences and allow for semantic similarity that captures, e.g., similarity between the words *guitar* and *music* because the guitar is an instrument used for creating music. For modern data, resources like WordNet [56] or DBpedia [12], their hierarchies or explicit relations can be used. For historical data where such resources may not be adequate, semantic similarity must be found using a combination of contemporary resources and unsupervised information extraction techniques. Because a large focus is given to historical resources in culturomics, finding semantic similarity measures and techniques for key resource finding is a challenge of great importance and future work is to investigate the utility of continuous word representations with temporal variance.

7 Discussion

As culturomics moves toward deeper analysis of large-scale corpora, the demand on processing capability moves beyond what any single computer can deliver regardless of its mem-

ory size or processor speed. This is a particularly important point as we move toward knowledge-based culturomics where many processes need more than one pass through the data and are computationally heavy. With the big data era, a new set of frameworks and tools have emerged that simplify the deployment of a distributed computing framework. To handle these challenges, focus must be given to large-scale distributed processing, for example, on frameworks like Hadoop [86].

Much of the discussion in this paper has regarded historical collections and the type of noise that must be handled with regard to, e.g., OCR errors and language change. However, it is important to keep in mind that similar problems arise when dealing with modern data, in particular from user-generated sources. A great variety in the form of slang, abbreviations and variations in spelling, word usage and grammar are present that can affect, e.g., NLP and semantic role labeling. To succeed with knowledge-based culturomics, attention must be paid to handle all these kinds of variation and provide methods that are robust against noise regardless of its origin.

One great advantage of knowledge-based culturomics is the possibility to iteratively apply different technologies and allow for automatic or semi-automatic improvements. As an example, while clustering word senses, we find that spelling and OCR errors often end up in the same clusters. Hence, using word sense clustering we can automatically determine which words to correct by, e.g., correcting words with low Levenshtein distance [45] to the word with the most frequent spelling. These corrections can be applied to the entire dataset and then clustering can take place again leading to higher quality clusters and possibly more or larger clusters. A concrete example is the following cluster derived from the Swedish Kubhist diachronic news corpus [80], for the year 1908: (*carbolinenm*, *kalk*, *cement*, *carbolineam*, *eldfast*, *carbolineum*). Three different spellings of *carbolineum* are present in the same cluster, two of them OCR errors which now can be corrected.

The strength of classical culturomics lies in its simplicity and low computational requirements. Because of the amount of data involved, many errors become statistically irrelevant and interesting patterns appear in the combination of (a) amount of data and (b) large timespans. However, there are many questions that cannot be answered using traditional culturomics; following concepts or entities over time that have changed their names or lexical representations, or distinguishing between different entities with the same name. Adding sentiment analysis, detecting changes in sentiment and the events that caused the changes is one example of moving beyond traditional culturomics.

Still, a deeper understanding of culturomics data requires human analysis. With knowledge-based culturomics the relevant resources can be found to help users focus their attention. Research questions like "what were the most relevant aspects of Marie Antoinette's life in terms of people, places, relations, what was the general sentiment of her and how and why did it change" can be answered. For each of these questions, documents can be provided for deeper analysis, going far beyond information retrieval with term matching techniques.

Current methods for NLP and statistical culturomics typically have the implicit assumption that the information expressed in documents is correct and can be taken at face value. For example, in opinion mining it is often assumed that the opinion expressed is the actual opinion of the speaker. First steps have been done to detected irony and sarcasm, where the implicit assumption obviously does not hold. However, this does not cover cases where authors misrepresent their opinions or factual knowledge for personal, political or social reason, e.g., to gain a financial advantage when talking about company stocks they own or to improve their standing in their peer group. Knowledge-based culturomics can be a first step to challenge the implicit assumption in two different ways: first, the richer model of the connections between entities, authors and documents makes it possible to aggregate first-order information in ways that are more representative than simple statistical summation. Furthermore, the findings of knowledge-based culturomics methods are linked back to source documents using key resource finding. This helps researchers validate the results instead of having to blindly trust the algorithms.

An interesting development in natural language processing, that we predict will have a strong impact on culturomics in the future, is the movement toward continuous vector space representation of words. These models embed words in a vector space that reveals semantic and syntactic similarities and can be used to lift the concrete text to a higher level of abstraction. This property was leveraged by Kågebäck et al. [36] to enhance the state-of-art extractive summarization algorithm introduced by Lin and Bilmes [47], using word embeddings to compare the information content in sentences. Further, these representations have been shown to exhibit interesting compositional properties, e.g., senses such as plurality and gender are captured as a linear translations in vector space, which can be exploited to draw conclusions not directly apparent from the text. An example of this is analogy testing, Mikolov et al. [55], where the question A relates to B as C relates to? Is answered using simple vector arithmetics $(v_B - v_A) + v_C \approx v_D$ where v_i denotes a vector representation of word i and v_D the vector representation of the sought answer. In culturomics word embeddings could be used for a variety of tasks, significantly extending the scope and impact of the project developed by Aiden and Michel [3] via the use of much richer representations in place of skip-grams. As a first example, we could directly compare words from different time periods by computing word embeddings

on time slices of history and projecting them into the same space, revealing trajectories for each word that plot the evolution of each corresponding word in a semantic space. Another example is to use the compositionality property, and compute how the vector space captures the popularity of a countries leader and subsequently track the evolution of a leaders corresponding word embedding for signs of dropping popularity, providing an early warning for civil unrest.

The aggregated data show that results of knowledge-based culturomics can be interesting for the general public and can help raise interest for historical resources. Imagine connecting modern newspaper articles with historical archives and for interesting current events being able to present the historical view. As an example, in an article on the topic *the Thailand Army declares martial law after unrest*, also historical aspects leading up to the event can be presented to users in condensed formats. Important people, places, events and different sentiments in summarized formats with links to descriptive original articles are all important to complement the current information and provide a more complete view.

8 Conclusion and future work

The notion of culturomics has turned out to be useful and interesting, not only from a research perspective but also for raising interest in historical data. However, without links to modern data it is less likely that culturomics will gain much interest from the general public. By moving toward knowledge-based culturomics, many more possibilities will be opened.

By offering support for integrating textual sources with modern and historical language, we can better bring information from historical resources (and the resources themselves) to users. This also allows researchers to better track information over long periods of time, without language change and variation getting in the way.

To achieve the goals set in this paper, particular focus must be given to natural language processing techniques and resources for entity, event and role extraction that constitute the basis of knowledge-based culturomics. Using this first-order information, knowledge-based culturomics opens up a large toolbox for researchers and the general public alike. Opinion mining and temporal semantic summaries allow for quick overviews and easy understanding of large amounts of textual resources over long periods of time. This analysis allows us to not only track information over time, but also clearly understand the change itself. As an example, we can find changes in opinion toward an entity over time as well as indications of the reason for the change, e.g., a specific event.

Finally, by linking extracted, second-order information to the primary resources, knowledge-based culturomics can offer credibility and traceability to this type of digital humanities research. Users can be offered a semantic interpretation of the important aspects of their queries as well as key resources that can serve as the starting point for further research.

While the benefits of knowledge-based culturomics are great, the challenges that lie ahead are equally large. Natural language processing is inherently hard as it is based on one of our most complex systems. However, tackling natural language at different periods in time, taking into consideration changes to the language, significantly increases the difficulty of the task. The increasing difficulty that comes with longer timespans applies to most other technologies that have thus far mostly been applied to data from short time periods. The strength of knowledge-based culturomics is the interconnected manner with which these challenges will be tackled, where the output of one technology can feed into another to improve the results.

Acknowledgments The authors would like to acknowledge the project "Towards a knowledge-based culturomics" supported by a framework Grant from the Swedish Research Council (2012–2016; dnr 2012-5738). We would also like to express our gratitude to the Centre for Language Technology in Gothenburg, Sweden (CLT, ⟨http://clt.gu.se⟩) for partial support. This work is also in parts funded by the European Commission under Alexandria (ERC 339233).

References

1. Adesam, Y., Ahlberg, M., Bouma, G.: bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa... towards lexical link-up for a corpus of Old Swedish. In: Proceedings of the 11th Conference on Natural Language Processing (KONVENS), Vienna, pp. 365–369. ÖGAI (2012). http://www.oegai.at/konvens2012/proceedings/54_adesam12w/54_adesam12w.pdf
2. Ahlberg, M., Bouma, G.: A best-first anagram hashing filter for approximate string matching with generalized edit distance. In: Proceedings of COLING 2012, Mumbai, pp. 13–22. ACL (2012). http://gup.ub.gu.se/records/fulltext/172769/172769.pdf
3. Aiden, E., Michel, J.-B.: Uncharted: Big Data as a Lens on Human Culture. Riverhead Books, New York (2013)
4. Allan, J., Gupta, R., Khandelwal, V.: Temporal summaries of new topics. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2001, pp. 10–18 (2001). doi:10.1145/383952.383954
5. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: Conference on Web Search and Data Mining, WSDM 2011, pp. 65–74 (2011). doi:10.1145/1935826.1935845
6. Balog, K., Mishne, G., de Rijke, M.: Why are they excited?: Identifying and explaining spikes in blog mood levels. In: Conference of the European Chapter of the Association for Computational

Linguistics: Posters & Demonstrations, EACL '06, pp. 207–210 (2006). http://dl.acm.org/citation.cfm?id=1608974.1609010
7. Bellare, K., McCallum, A.: Learning extractors from unlabeled text using relevant databases. In: Sixth International Workshop on Information Integration on the Web (2007)
8. Berberich, K., Bedathur, S.J., Sozio, M., Weikum, G.: Bridging the terminology gap in web archive search. In: Proceedings of the 12th International Workshop on the Web and Databases, WebDB 2009 (2009). http://webdb09.cse.buffalo.edu/papers/Paper20/webdb2009-final.pdf
9. Bhattacharya, I., Getoor, L.: A latent Dirichlet model for unsupervised entity resolution. In: Siam International Conference on Data Mining (2006)
10. Bhattacharya, I., Getoor, L.: Collective entity resolution in relational data. ACM Trans. Knowl. Discov. Data (TKDD) **1**(1) (2007). doi:10.1145/1217299.1217304
11. Binh Tran, G.: Structured summarization for news events. In: International Conference on World Wide Web Companion, WWW '13 Companion, pp. 343–348 (2013). http://dl.acm.org/citation.cfm?id=2487788.2487940
12. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia—a crystallization point for the Web of Data. J. Semant. **7**(3), 154–165 (2009). doi:10.1016/j.websem.2009.07.002
13. Bunescu, R.C., Mooney, R.: Learning to extract relations from the web using minimal supervision. In: Annual Meeting of the Association for Computational Linguistics, ACL 2007, p. 576 (2007)
14. Calais Guerra, P.H., Veloso, A., Meira Jr, W., Almeida, V.: From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: Conference on Knowledge Discovery and Data Mining, KDD 2011, pp. 150–158 (2011). doi:10.1145/2020408.2020438
15. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.: Measuring user influence in twitter: The million follower fallacy. In: International AAAI Conference on Weblogs and Social Media, ICWSM 2010 (2010). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1538
16. Choi, Y., Breck, E., Cardie, C.: Joint extraction of entities and relations for opinion recognition. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2006, pp. 431–439 (2006)
17. Craven, M., Kumlien, J., et al.: Constructing biological knowledge bases by extracting information from text sources. In: Conference on Intelligent Systems for Molecular Biology, pp. 77–86 (1999)
18. Demartini, G., Siersdorfer, S., Chelaru, S., Nejdl, W.: Analyzing political trends in the blogosphere. In: Fifth International AAAI Conference on Weblogs and Social Media, ICWSM 2011 (2011)
19. Ernst-Gerlach, A., Fuhr, N.: Retrieval in text collections with historic spelling using linguistic and spelling variants. In: Joint International Conference on Digital Libraries, JCDL 2007, pp. 333–341 (2007). doi:10.1145/1255175.1255242
20. Exner, P., Nugues, P.: Constructing large proposition databases. In: International Conference on Language Resources and Evaluation, LREC 2012, p. 5 (2012)
21. Filatova, E., Hatzivassiloglou, V.: A formal model for information selection in multi-sentence text extraction. In: International Conference on Computational Linguistics, COLING 2004 (2004). doi:10.3115/1220355.1220412
22. Fillmore, C.J.: Frame semantics and the nature of language. Ann. N. Y. Acad. Sci. **280**, 20–32 (1976)
23. Ganesan, K., Zhai, C., Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: International Conference on Computational Linguistics, COLING 2010, pp. 340–348 (2010). http://dl.acm.org/citation.cfm?id=1873781.1873820
24. Gillick, D., Favre, B., Hakkani-tür, D., Bohnet, B., Liu, Y., Xie, S.: The ICSI/UTD summarization system at TAC 2009. In: Text Analysis Conference (2009)
25. Google Books. http://books.google.com/ (2013). Retrieved 26 June 2013
26. Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., Schulz, K.U.: Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In: Workshop on Analytics for Noisy Unstructured Text Data, AND 2009, pp. 69–76 (2009). doi:10.1145/1568296.1568309
27. Greene, S., Resnik, P.: More than words: syntactic packaging and implicit sentiment. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, pp. 503–511. ACLs (2009). http://www.aclweb.org/anthology/N/N09/N09-1057
28. Günther, T.: Sentiment analysis of microblogs. Master's thesis, University of Gothenburg (2013)
29. Hachey, B.: Multi-document summarisation using generic relation extraction. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, pp. 420–429 (2009). http://dl.acm.org/citation.cfm?id=1699510.1699565
30. Haghighi, A., Klein, D.: Coreference resolution in a modular, entity-centered model. In: Human Language Technologies, HLT 2010, pp. 385–393 (2010). http://dl.acm.org/citation.cfm?id=1857999.1858060
31. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Annual Meeting of the Association for Computational Linguistics and Conference of the European Chapter of the Association for Computational Linguistics, pp. 174–181 (1997)
32. Hauser, A., Heller, M., Leiss, E., Schulz, K.U., Wanzeck, C.: Information access to historical documents from the Early New High German Period. In: Digital Historical Corpora—Architecture, Annotation, and Retrieval, number 06491 in Dagstuhl Seminar Proceedings (2007). http://drops.dagstuhl.de/opus/volltexte/2007/1057
33. Hermansson, L., Kerola, T., Johansson, F., Jethava, V., Dubhashi, D.: Entity disambiguation in anonymized graphs using graph kernels. In: International Conference on Information and Knowledge Management, CIKM '13, pp. 1037–1046 (2013). doi:10.1145/2505515.2505565
34. Ji, H., Favre, B., Lin, W.-P., Gillick, D., Hakkani-Tur, D., Grishman, R.: Open-domain Multi-Document summarization via information extraction: Challenges and prospects. In: Saggion, H., Poibeau, T., Yangarber, R. (eds.) Multi-source Multilingual Information Extraction and Summarization. Lecture Notes in Computer Science. Springer (2011)
35. Johansson, R., Alessandro, M.: Relational features in fine-grained opinion analysis. Comput. Linguist. **39**(3), 473–509 (2013)
36. Kågebäck, M., Mogren, O., Tahmasebi, N., Dubhashi, D.: Extractive summarization using continuous vector space models. In: Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), Gothenburg, Sweden, pp. 31–39. Association for Computational Linguistics (2014). http://www.aclweb.org/anthology/W14-1504
37. Kaluarachchi, A., Roychoudhury, D., Varde, A.S., Weikum, G.: SITAC: discovering semantically identical temporally altering concepts in text archives. In: International Conference on Extending Database Technology, EDBT/ICDT '11, pp. 566–569 (2011). doi:10.1145/1951365.1951442
38. Kawahara, D., Shinzato, K., Shibata, T., Kurohashi, S.: Precise information retrieval exploiting predicate–argument structures. In: Proceeding of the IJCNLP (2013)

39. Kim, S.-M., Hovy, E.: Extracting opinions, opinion holders, and topics expressed in online news media text. In: Workshop on Sentiment and Subjectivity in Text, pp. 1–8 (2006)
40. Lau, J.H., Cook, P., McCarthy, D., Newman, D., Baldwin, T.: Word sense induction for novel sense detection. In: Conference of the European Chapter of the Association for Computational Linguistics, EACL 2012, pp. 591–601 (2012). http://aclweb.org/anthology-new/E/E12/E12-1060.pdf
41. Lazaridou, A., Titov, I., Sporleder, C.: A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In: Annual Meeting of the Association for Computational Linguistics, ACL 2013, pp. 1630–1639 (2013)
42. Lenhert, W., Cardie, C., Fisher, D., Riloff, E., Williams, R.: Description of the CIRCUS system as used for MUC-3. In: Message Understanding Conference. Morgan Kaufmann (1991). http://acl.ldc.upenn.edu/M/M91/M91-1033.pdf
43. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 497–506 (2009). doi:10.1145/1557019.1557077
44. Leskovec, J., Grobelnik, M., Milic-Frayling, N.: Learning substructures of document semantic graphs for document summarization. In: Workshop on Link Analysis and Group Detection, LinkKDD 2004 (2004)
45. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. **10**(8), 707–710 (1966)
46. Liberman, M.: String frequency distributions. In: Language Log posting, 3rd Feb (2013). http://languagelog.ldc.upenn.edu/nll/?p=4456
47. Lin, H., Bilmes, J.: A class of submodular functions for document summarization. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL 2011, pp. 510–520 (2011)
48. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Conference on Information and Knowledge Management, CIKM 2009, pp. 375–384 (2009)
49. Lin, C.-H., Yen, C.-W., Hong, J.-S., Cruz-Lara, S., et al.: Event-based textual document retrieval by using semantic role labeling and coreference resolution. In: IADIS International Conference WWW/Internet 2007 (2007)
50. Liu, B.: Sentiment analysis and opinion mining. In: Synthesis Lectures on Human Language Technologies. Morhan & Claypool Publishers (2012)
51. Mann, J., Zhang, D., Yang, L., Das, D., Petrov, S.: Enhanced search with wildcards and morphological inflections in the Google Books Ngram Viewer. In: Proceedings of ACL Demonstrations Track, Baltimore. ACL (2014) (to appear)
52. McCarthy, J.F., Lehnert, W.G.: Using decision trees for coreference resolution. In: International Joint Conference On Artificial Intelligence, pp. 1050–1055 (1995)
53. Michel, J.-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011)
54. Mihalcea, R., Tarau, P.: Textrank: Bringing order into texts. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 (2004)
55. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746–751 (2013). http://www.aclweb.org/anthology/N13-1090
56. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**, 39–41 (1995)
57. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Joint Conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP, ACL 2009, pp. 1003–1011 (2009)
58. Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., Goyal, P.: That's sick dude!: Automatic identification of word sense change across different timescales. CoRR, abs/1405.4392 (2014). http://arxiv.org/abs/1405.4392
59. Moretti, F.: Graphs, Maps, Trees: Abstract Models for a Literary History. Verso (2005). ISBN 9781844670260
60. Newcombe, H.B., Kennedy, J.M., Axford, S.J., James, A.P.: Automatic linkage of vital records. Science **130**(3381), 954–959 (1959)
61. Ng, V.: Unsupervised models for coreference resolution. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, pp. 640–649 (2008)
62. Nguyen, T., Phung, D., Adams, B., Venkatesh, S.: Event extraction using behaviors of sentiment signals and burst structure in social media. Knowl. Inf. Syst. 1–26 (2012)
63. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. In: International AAAI Conference on Weblogs and Social Media, ICWSM 2010 (2010)
64. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania, United States, pp. 79–86 (2002). doi:10.3115/1118693.1118704
65. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
66. Poon, H., Domingos, P.: Joint unsupervised coreference resolution with markov logic. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, pp. 650–659 (2008). http://www.aclweb.org/anthology/D08-1068
67. Project Gutenberg. http://www.gutenberg.org/. (2013). Retrieved 26 June 2013
68. Radev, D.R., Jing, H., Styś, M., Tam, D.: Centroid-based summarization of multiple documents. Inf. Process. Manag. **40**(6), 919–938 (2004)
69. Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., Manning, C.D.: A multi-pass sieve for coreference resolution. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 492–501 (2010)
70. Rahman, A., Ng, V.: Coreference resolution with world knowledge. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT 2011, pp. 814–824 (2011). http://dl.acm.org/citation.cfm?id=2002472.2002575
71. Rastogi, V., Dalvi, N., Garofalakis, M.: Large-scale collective entity matching. VLDB Endow. **4**(4), 208–218 (2011). http://dl.acm.org/citation.cfm?id=1938545.1938546
72. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Machine Learning and Knowledge Discovery in Databases, vol. 6323 of LNCS, pp. 148–163. Springer (2010)
73. Ruppenhofer, J., Somasundaran, S., Wiebe, J.: Finding the sources and targets of subjective expressions. In: International Conference on Language Resources and Evaluation, LREC 2008, pp. 2781–2788 (2008)
74. Rusu, D., Fortuna, B., Grobelnik, M., Mladenic, D.: Semantic graphs derived from triplets with application in document summarization. Informatica (Slovenia) **33**(3), 357–362 (2009)
75. Sagi, E., Kaufmann, S., Clark, B.: Semantic density analysis: comparing word meaning across time and phonetic space. In: Workshop on Geometrical Models of Natural Language Semantics, GEMS 2009, pp. 104–111 (2009). http://dl.acm.org/citation.cfm?id=1705415.1705429

76. Somasundaran, S., Namata, G., Wiebe, J., Getoor, L.: Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, Singapore, pp. 170–179 (2009)
77. Tahmasebi, N., Gossen, G., Kanhabua, N., Holzmann, H., Risse, T.: NEER: an unsupervised method for Named Entity Evolution Recognition. In: International Conference on Computational Linguistics, COLING 2012, pp. 2553–2568 (2012). http://www.aclweb.org/anthology/C12-1156
78. Tahmasebi, N.: Models and algorithms for automatic detection of language evolution. Ph.D. thesis, Gottfried Wilhelm Leibniz Universität Hannover (2013)
79. Tahmasebi, N., Niklas, K., Zenz, G., Risse, T.: On the applicability of word sense discrimination on 201 years of modern english. Int. J. Digit. Libr. **13**(3–4), 135–153 (2013). doi:10.1007/s00799-013-0105-8. ISSN 1432-5012
80. The Kubhist Corpus. http://spraakbanken.gu.se/korp/?mode=kubhist. Språkbanken, Department of Swedish, University of Gothenburg
81. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in twitter events. J. Am. Soc. Inf. Sci. Technol. **62**(2), 406–418 (2011). doi:10.1002/asi.21462
82. Tsytsarau, M., Palpanas, T.: Survey on mining subjective data on the web. Data Min. Knowl. Discov. **24**, 478–514 (2012). doi:10.1007/s10618-011-0238-6
83. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Annual Meeting of the Association for Computational Linguistics, ACL 2002, pp. 417–424 (2002)
84. Velikovich, L., Blair-Goldensohn, S., Hannan, K., McDonald, R.: The viability of web-derived polarity lexicons. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL 2010, pp. 777–785 (2010)
85. Whang, S.E., Garcia-Molina, H.: Entity resolution with evolving rules. VLDB Endow. **3**(1–2), 1326–1337 (2010). http://dl.acm.org/citation.cfm?id=1920841.1921004
86. White, T.: Hadoop: The Definitive Guide. O'Reilly Media Inc (2012)
87. Wiebe, J., Bruce, R., O'Hara, T.: Development and use of a gold standard data set for subjectivity classifications. In: Annual Meeting of the Association for Computational Linguistics, ACL 1999, pp. 246–253 (1999)
88. Wijaya, D.T., Yeniterzi, R.: Understanding semantic change of words over centuries. In: Workshop on DETecting and Exploiting Cultural diversiTy on the social web, DETECT 2011, pp. 35–40 (2011). doi:10.1145/2064448.2064475
89. Wilson, T.A.: Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states. Ph.D. thesis, University of Pittsburgh, Pittsburgh, United States (2008)
90. Wu, Y., Oard, D.W.: Beyond topicality, finding opinionated documents. In: Annual Conference of the Association for Information Science and Technology, Vancouver (2000)
91. Wu, F., Weld, D.S.: Autonomously semantifying Wikipedia. In: Conference on Information and Knowledge Management, CIKM 2007, pp. 41–50 (2007)
92. Yan, R., Kong, L., Huang, C., Wan, X., Li, X., Zhang, Y.: Timeline generation through evolutionary trans-temporal summarization. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, pp. 433–443 (2011). http://dl.acm.org/citation.cfm?id=2145432.2145483
93. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2003, pp. 129–136 (2003)