Mining semantics for culturomics: towards a knowledge-based approach

Lars, Borin; Devdatt, Dubhashi; Markus, Forsberg; Johansson, Richard; Dimitrios, Kokkinakis; Nugues, Pierre

Published in:
UnstructureNLP '13 Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing

DOI:
10.1145/2513549.2513551

2013

Mining Semantics for Culturomics – Towards a Knowledge-Based Approach

Lars Borin
Språkbanken, Dept. of Swedish
University of Gothenburg
SE-405 30, Gothenburg
Sweden
lars.borin@gu.se

Devdatt Dubhashi
Computer Science & Engineering
Chalmers University of Technology
SE-412 96, Gothenburg
Sweden
dubhashi@chalmers.se

Markus Forsberg
Språkbanken, Dept. of Swedish
University of Gothenburg
SE-405 30, Gothenburg
Sweden
markus.forsberg@gu.se

Richard Johansson
Språkbanken, Dept. of Swedish
University of Gothenburg
SE-405 30, Gothenburg
Sweden
richard.johansson@gu.se

Dimitrios Kokkinakis
Språkbanken, Dept. of Swedish
University of Gothenburg
SE-405 30, Gothenburg
Sweden
dimitrios.kokkinakis@gu.se

Pierre Nugues
Department of Computer science
Lund University
SE-223 63, Lund
Sweden
pierre.nugues@cs.lth.se

ABSTRACT

The massive amounts of text data made available through the Google Books digitization project have inspired a new field of big-data textual research. Named *culturomics*, this field has attracted the attention of a growing number of scholars over recent years. However, initial studies based on these data have been criticized for not referring to relevant work in linguistics and language technology. This paper provides some ideas, thoughts and first steps towards a new culturomics initiative, based this time on Swedish data, which pursues a more knowledge-based approach than previous work in this emerging field. The amount of new Swedish text produced daily and older texts being digitized in cultural heritage projects grows at an accelerating rate. These volumes of text being available in digital form have grown far beyond the capacity of human readers, leaving automated semantic processing of the texts as the only realistic option for accessing and using the information contained in them. The aim of our recently initiated research program is to advance the state of the art in language technology resources and methods for semantic processing of *Big Swedish text* and focus on the theoretical and methodological advancement of the state of the art in extracting and correlating information from large volumes of Swedish text using a combination of knowledge-based and statistical methods.

Categories and Subject Descriptors

H.3.1 **[Content Analysis and Indexing]**; H.3.3 **[Information Search and Retrieval]**; H.3.7 **[Digital Libraries]**; I.2.4 **[Knowledge Representation Formalisms and Methods]**; I.2.7 **[Natural Language Processing]**

UnstructureNLP'13, October 28, 2013, San Francisco, CA, USA.
ACM 978-1-4503-2415-1/13/10.
http://dx.doi.org/10.1145/2513549.2513551.

General Terms

Algorithms, Experimentation, Management, Standardization.

Keywords

culturomics, semantics, language processing, Big data, Swedish.

1. INTRODUCTION

The main aim of our recently initiated research program, *Knowledge-based culturomics*, is to advance the state of the art in language technology resources and methods for deep linguistic processing of Swedish text. In this way, we can provide researchers in several fields with more sophisticated tools for working with the information contained in the large volumes of digitized text, e.g., by being able to correlate and compare the content of texts and text passages on a large scale. The project focus is on the design, integration and adaptation of language-processing components for semantic analysis (e.g., entities, relations, events and their structure, their semantic roles, and coreference between their arguments). The results will provide researchers with more sophisticated tools for working with the information contained in large volumes of digitized text and develop methodology and applications in support of research in disciplines where text is an important primary research data source, primarily the humanities and social sciences, but which also will benefit everyone else who works within the (Swedish) language landscape.

2. CULTUROMICS

Culturomics is an emerging new research area born of the combination of large-scale digitization and basic language processing. It started out as a methodology for studying culture and language development over time using the massive Google Books dataset [44, 50]. With the availability of big unstructured text data [49], such as in the form of millions of digitized books or other large collections of text (and data) as in the healthcare sector [32], the possibilities of culturomics are almost limitless. The practice of culturomics, as presented by the originators of the concept, is inspired by work originating in computer science,

especially information retrieval, where the input data generally is made up by linguistically unanalyzed text units – "bags of words" – but where we occasionally also see some attempts to benefit from deeper linguistic processing, for example *topic models* [7, 30]. The main objective of culturomics is to assemble and reorganize data in order to offer researchers from various disciplines a new playground to investigate quantitatively, e.g., broad cultural trends and the cultural developments of various societies through time. Culturomics also relies on statistically-oriented techniques in order to track long-term, macro-scale patterns of written language use and derive conclusions about language that have never previously been possible in the study of human language. For instance, Acerbi et al. [1] analyze trends in the past century of mood words (emotion-related words) in books using the Google n-gram database.

Michel et al. [44] mention that big data "will furnish a great cache of bones from which to reconstruct the skeleton of a new science." and that "there are strong parallels to the completion of the human genome (analogously to genomics, proteomics and other -omic technologies). Just as that provided an invaluable resource for biologists, Google's corpus will allow social scientists and humanities scholars to study human culture in a rigorous way". Note, that in a wider sense the goal of culturomics is related to the idea of *macroanalysis* by Moretti [46] who coined the term "distant reading" in which "the reality of the text undergoes a process of deliberate reduction and abstraction". According to this view, understanding language (e.g., literature) is not accomplished by studying individual texts, but by aggregating and analyzing massive amounts of data [34]. In this way it becomes possible to design experiments for investigating novel uses of language and its development that otherwise would be impossible to conduct, e.g., quantifying the difference between writing styles [3].

We are certainly about to see new developments in the exploration of cultural trends in the near future by the use of new enhanced culturomics initiatives applied on big data worldwide. *Twitterology*, a subfield of culturomics, is such a development that focuses on large tweets corpora [54] and *Culturomics 2.0* proposed by e.g., Leetaru [40], a new development that can be used to defuse some of the criticisms raised against culturomics [31]. Moreover, Goldberg and Orwant [28] have recently created and released a dataset of syntactic n-grams based on the Google Books corpus, which contains over 10 billion unique items covering various syntactic structures, including temporal metadata which can facilitate research into lexical semantics over time.

3. RESOURCES

This project deals with written Swedish language using available digital text collections and corpora. Therefore, we strongly believe that the basic premise of the culturomics endeavor is eminently timely. Large amounts of new Swedish text are produced constantly while older texts are being digitized apace in cultural heritage projects – e.g., in the Digidaily[1] project and the Swedish Literature Bank.[2] The volumes of Swedish text available in digital form have grown far beyond the capacity of even the fastest reader, leaving automated semantic processing of the texts, such as production of semantically-oriented structured data and feature/value set extraction, as the only realistic option for accessing and using this information. However, advanced technologies that require semantic processing are still very much in their infancy for almost all languages, Swedish being no exception. The data we started using comes from the University of Gothenburg's Swedish Language Bank (Språkbanken, SB[3]). SB maintains, develops and annotates a corpus collection of over one billion words of modern texts and close to one billion words of historical texts. SB's corpus collection already contains the full Swedish Wikipedia and about 390 MW of Swedish blogs. Collecting large amounts of text from the web is of course not a technical problem [3], but it does raise interesting issues of how to work effectively with very large volumes of text.

4. LANGUAGE TECHNOLOGY AND SEMANTIC PROCESSING

Natural language is not simply a set of invariant words. It has linear and hierarchical structure, and the data encountered in different databases, archives and real time, social media texts, is "noisy" thanks to, e.g., variations in spelling, capitalization, inflection, and ambiguities of various kinds. This project has started using semantically-oriented tools and resources, a brief description of which is provided below, in order to effectively deal with many of the linguistic issues encountered in raw data.

4.1 Entities and Coreference

Named entity recognition (NER) is a core subtask in Information Extraction and has emerged as an important supporting component with many applications in various human language technologies. The automatic recognition and marking-up of names (in a wide sense) and various other related kinds of information, e.g., time, measure expressions and/or terminology, has turned out to be a recurring basic requirement. Hence, NER has become a core language technology of great significance to numerous applications and a wide range of techniques. However, the nature and type of named entities vary, depending on the task under investigation or the target application. In any case, person, location and organization names are considered "generic", in the sense that all NER systems incorporate a mechanism to identify these basic entities. In this work we apply a system using rather fine grained named entity taxonomy with several main entity types and subtypes. For instance, the type person incorporates peoples' names (forenames, surnames), groups of people, animal/pet names, mythological names, theonyms and the like, for peoples' names, even gender identification is performed. In previous studies with the same system, acceptable figures on precision and recall have been reported [12, 13]. Discourse entities are the real, abstract, or imaginary objects introduced in a text. A correct detection and identification of the entities

[1] Digidaily is a project co-financed by the European Union's Structural Funds. The Swedish Royal Library and the National Archives have to date digitized almost 300,000 pages of Swedish newspaper text from the 17th century onwards.

[2] The Swedish Literature Bank contains over 100,000 pages of classical Swedish literature in digital text editions. The aim of the Swedish Literature Bank is to be a free cultural-historical and literary resource for research, teaching and popular education. The main task is to collect and digitize literary works and the most important works of the humanities, and to make the material available in a way that makes it possible for users to work with it.

[3] For a more detailed description of the linguistic infrastructure in Språkbanken visit: <http://spraakbanken.gu.se/> moreover, SB's constantly evolving corpus search interface [11] is open for general use here: <http://spraakbanken.gu.se/korp>.

mentioned in a text is crucial for its understanding. As entities may be mentioned two or more times, a coreference solver is needed to link the expressions that refer to the same entity [6, 48]. A coreference solver builds sets of equivalence classes referring to these mentions in the text, e.g., linking an anaphor and its antecedent, as being references to the same real-world entity or object. Coreference solving is considered complex and sometimes a show-stopper in semantic applications.

4.2 Relations and Events

There are two basic ways to structure and further explore text data in textual resources of various kinds. Apart from annotating the unstructured text with entities we also need to identify relations both between entities and between events. Semantic relations are a connection between combinations of things (here named entities) in text. Semantic relations are usually binary, and can be expressed as grammaticosemantic links between lexical units, e.g. named entities, for example: *X father-of Y* or *father-of(X, Y)* where X and Y are instances of named entity types; by syntactic dependencies, which approximate the underlying semantic relationships [15]; or by conventional RDF S-V-O triples, which is a universal representation of relations. Moreover, semantic relations are often terminological relations, e.g. *Part-of* or *Is-a* or static relations such as *Born-In*. On the other hand, events, (see below), are dynamic.

In this work we do not differentiate between these different types but let the data "decide" what relations and events are present. Events (in linguistics usually defined in short as "things that happen") are structurally far more complex processes than relations and often are expressed in natural language with the help of predicates (e.g., verbs) that assign event-specific semantic roles to each participating entity, linking them to (syntactic) arguments. These relations between event predicates and arguments (predicate-argument relations) are the focus of semantic role labeling methods that automatically assign roles to arguments of a predicate. An event is exemplified with the help of the sample sentence provided in Figure 1: *Öroninflammation (akut otitis media, AOM) är en vanlig barnsjukdom och den vanligaste orsaken till att barn erhåller antibiotikabehandling.* 'Ear infections (acute otitis media, AOM) is a common childhood disease and the most common reason that children receive antibiotic treatment.'. This sentence contains several events, for instance a *Causation* event, triggered by the predicate *orsaken*, 'the reason', and a *Medical-Treatment* event, triggered by the predicate *erhåller* '(they) receive'. For event recognition we have started experimentation with the Swedish FrameNet++ (Section 4.3); therefore these two event names are instantiated using two frames with the same name in FrameNet++.

4.3 Frame Semantics and FrameNet

The FrameNet approach is based on the linguistic theory of frame semantics [25] supported by corpus evidence. A semantic frame is a script-like structure of concepts which are linked to the meanings of linguistic units and associated with a specific event or state. Each frame identifies a set of frame elements, which are frame specific semantic roles (both core and non-core ones).

Furthermore, roles may be expressed overtly, left unexpressed or not explicitly linked to the frame via linguistic conventions. In this work, we only deal with the first type of such roles. Semantic lexicons, such as FrameNet, are an important source of knowledge for semantic parsing (Section 4.4). Despite the belief that certain Natural Language Processing (NLP) tasks, such as statistical machine translation and speech recognition, could be performed more efficiently and less expensively by pure statistical means [29], FrameNet relies on expert annotations and much less data than e.g., Google's n-gram database (sequences of *n* words), in order to capture semantic subtleties that cannot be captured by statistical means. Therefore, we believe that "big data" will need NLP, but the NLP also in its turn needs "big data", since current approaches based merely on e.g., n-gram modeling and collocation scores are hardly more than advanced means of conducting naïve searches and a new way of performing frequency analysis. On the other hand, FrameNet documents the range of semantic and syntactic combinatory possibilities of frame evoking lexical units (LU), phrases and clauses by abstracting away from syntactic differences, which can be beneficial for a deeper exploration of language. An LU can evoke a frame, and its syntactic dependents can fill the frame element slots (see the annotations in Fig. 1). The Swedish FrameNet++ [10], SweFN++, is a lexical resource under active development in SB, based on the English version of FrameNet. It is found on the SweFN website,[4] available as a free resource under a CC BY license.

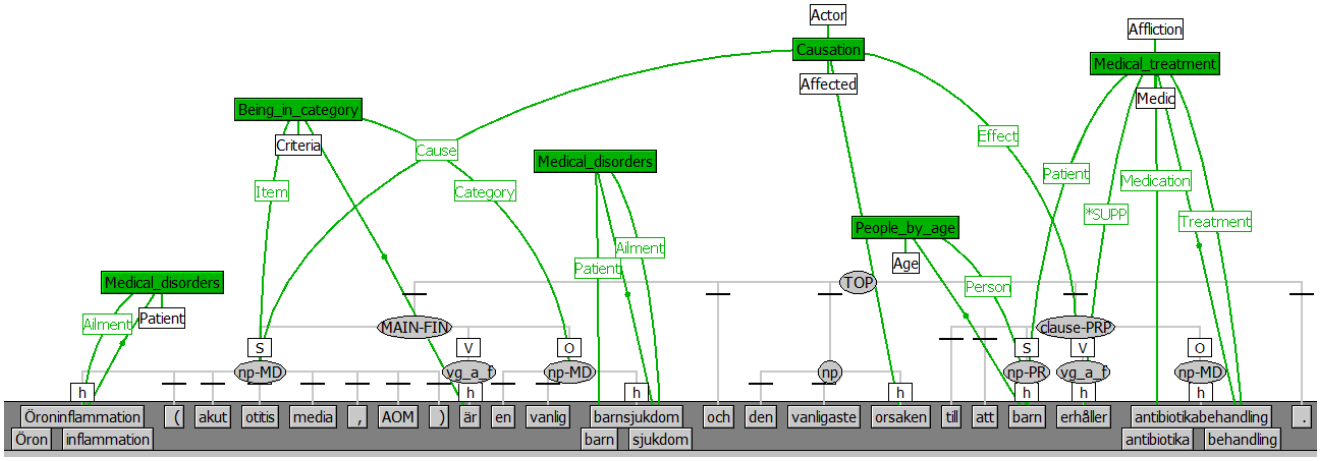

Figure 1. Annotated sentence in SALTO [14], using a combination of shallow syntax and frame semantics. Here, gray labels are the result from the syntactic analysis and the green boxes are frames and (a subset of) their elements.

[4] http://spraakbanken.gu.se/eng/swefn.

4.4 Semantic Parsing

With the advent of massive online encyclopedic corpora such as Wikipedia, it has become possible to apply a systematic analysis to a wide range of documents covering a significant part of human knowledge. Semantic parsers or related techniques then enable machines to extract such knowledge in the form of propositions (predicate–argument structures) and build large proposition databases from these documents. Therefore, semantic parsing is understood as the transformation of natural language sentences into complete computer-executable meaning representations for domain-specific applications. A shallow form of such semantic representation is a semantic role labeling, which, according to the underlying theoretical model, identifies roles such as *Agent*, *Purpose*, *Theme* and *Instrument*.

Systems like IBM Watson [22] and OLLIE [42] have carried out a systematic extraction of semantic roles or frames on very large corpora. Both systems used grammatical relations and rules to derive the predicate-argument structures, as this technique is fast and relatively easy to apply to large corpora. A statistical semantic role labeler is slower, but usually more accurate. Christensen et al. [17] showed that using a semantic parser in information extraction can yield a higher precision and recall. Nonetheless, it is possible to combine both techniques to get higher performances. See Mausam et al. [42] for a discussion.

Semantic analysis provides a representation of the sentence formally which can further support automated reasoning. We have started to explore learning semantic parsers for mapping natural-language sentence components to semantic roles, by applying the results of the ongoing effort to develop a Swedish FrameNet lexicon (see the previous section and [35]). However, in order to achieve a competitive performance for this system, a combination of rule-based techniques with lightly supervised or even unsupervised learning techniques may be necessary [26, 27, 51]. Moreover, since semantic parsing requires some form of preliminary syntactic analysis, the semantic parsers we envisage will be based on syntactic dependencies and/or shallow constituent structures (Figure 1).

4.5 Bootstrapping Analyzers

The base semantic parser we developed [36] uses supervised-learning techniques and requires annotated corpora up to the semantic level. For a language like Swedish, this might be problematic and impair the parsing quality as semantically-annotated corpora available for it are much smaller than their counterparts in English. To offset this lack of large corpora, we are exploring distant supervision techniques and the mapping of entities and relations in loosely parallel texts such as the Swedish and English versions of Wikipedia.

Distant supervision [45, 47] uses existing relation databases such as YAGO [53] and DBpedia [2, 5] and entities occurring in the relations. DBpedia is an RDF knowledge base containing facts extracted from Wikipedia semi-structured information, notably the infoboxes, small table summarizing an article. For each tuple of entities, distant supervision extracts all the sentences from large, unannotated corpora that contain the tuple. It is then possible to train new classifiers to identify the relations. We are investigating such techniques across languages, notably with the pair Swedish-English, in combination with dependency parsing. The goal is to link entities across Wikipedia versions to the YAGO and DBpedia entity knowledge bases and relations using grammatical functions and predicate-argument structures.

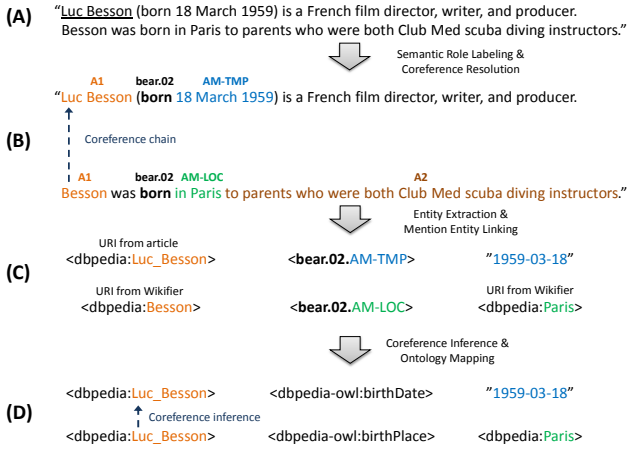

Figure 2. A conversion from text to DBpedia RDF triples: (A) Input sentences. (B) Sentences after semantic parsing and coreference resolution. (C) Entity extraction. (D) Ontology mapping. After Exner and Nugues [20].

Figure 2 shows an example of such mapping ideas restricted to one language, English for now. A first step extracts all the propositions from the complete collection of Wikipedia articles in English, here using the Athena system [18, 19], and solves the coreferences. A second step maps the extracted propositions to DBpedia facts. Figure 2 shows this mapping from PropBank predicates to DBpedia properties and entities. Finally, the pairs are used to map DBpedia properties to PropBank predicates, complement the RDF triple repository, and extend the annotation of the corpus [20, 21].

5. APPLICATION SCENARIOS

This section describes three application scenarios with different ambitions and goals we envisage in the project (some with very preliminary results), namely question answering, tracking of semantic change of word meaning including diachronic development of words, and document summarization using a knowledge based integrative approach. Naturally, all scenarios involve important issues hidden under the surface (e.g., a lot of low-level NLP tasks some of which are outlined in Section 4) that needs to be taken into consideration. Importantly, there is a large overlap between the resources and tool components needed for the three scenarios, due to the project focus on knowledge-based semantic processing of big text data.

Question answering systems are notable applications of semantic processing. They reached a milestone in 2011 when IBM's Watson DeepQA system [24] outperformed all its human co-contestants in the *Jeopardy!* quiz show [22, 23, 24]. Watson answers questions in any domain[5] posed in natural language using knowledge extracted from Wikipedia and other textual sources, encyclopedias, dictionaries such as WordNet, as well as databases

[5] Chaudhry [16] presents the deployment of IBM Watson in the oncology domain and utilization management. Chaudhry emphasizes Watson's capabilities such as "natural language understanding" and "iterative Q/A" as well a "shallower" reasoning over large volumes of data, could be applied in healthcare.

such as DBpedia and YAGO [53] (see Section 4.5). A goal of the project which will ensure its visibility is to replicate the IBM Watson system for Swedish with knowledge extracted from different sources. Searching in a semantically-oriented manner is one of the main priorities of our culturomics project.

The second application scenario the project intends to investigate is *tracking semantic change.* The problem of tracking semantic change in searching document archives has been addressed recently [4, 52], e.g., by Berberich et al. [4], who proposed a solution to this problem by reformulating a query into terms prevalent in the past by comparing the contexts over time-captured by co-occurrence statistics. The approach requires a recurrent computation which can affect efficiency and scalability. Kaluarachchi et al. [37, 38] proposed to discover semantically identical concepts (or named entities) that are used at different times using an association rule mining technique using events (sentences containing a subject, a verb, objects, and nouns) associated to two distinct entities. Two entities are semantically related if the associated events occur multiple times in a document archive. The approach relies on linguistic properties and events, which are subjected to change over time as well. Kanhabua and Nørvåg [39] tracked named entity changes from anchor texts in Wikipedia and associated each version of a term with a period of validity using Wikipedia history as well as New York Times Annotated Corpus. Unfortunately, the method has limited applicability as link information, such as anchor texts, is not always available in other document archives. In more recent work, Mazeika et al. [43] extracted named entities from the YAGO ontology and tracked their changed usage patterns using the New York Times Annotated Corpus. Similar to the work by Kanhabua and Nørvåg [39], relying on the ontological knowledge is expensive and requires human annotators.

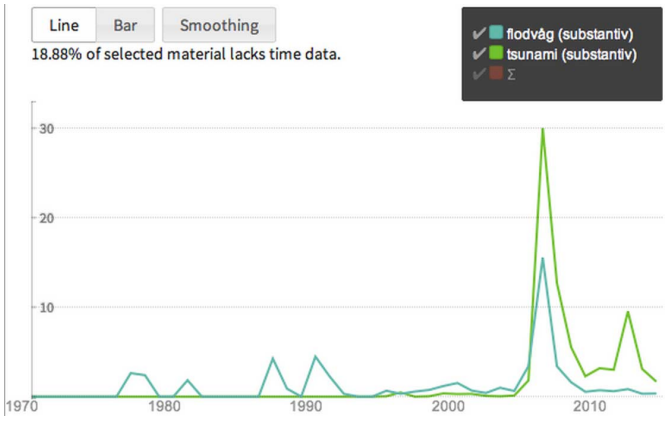

Figure 3. Exploring word usage over time.

Figure 3 shows a distribution graph for two Swedish near synonyms, namely *tsunami* and *flodvåg* 'tidal wave' (including all inflectional forms and all compounds containing these words) in Swedish news texts since the beginning of the 1970s. Here we can see that the word *flodvåg* has a longer history in the language and that its synonym *tsunami* gained a secure foothold only after the December 2004 Indian Ocean earthquake and tsunami event. This is a kind of result that emerge directly from a linguistically annotated text material which is made possible by the lexical analysis tools based on handcrafted resources used for annotating Språkbanken's corpora.

Finally, a third application scenario is document summarization. Since our collective knowledge (e.g., news, blogs, scientific articles, books etc.) continues to be digitized and stored it becomes more difficult to find and discover what we are looking for. Therefore we strongly believe that one tool to help us organize, search and understand this vast amount of information is to use automatic document summarization. The *extractive summarization* problems is: Given a document or set of documents consisting of a set of sentences V, extract a summary $S \subseteq V$ that is representative of document i.e. identify important/representative sentences or pieces of text which contain the information about the most important concepts mentioned therein. The problem is posed in a fully unsupervised setting i.e. there is no training data and it is required to be relatively domain and even language independent [9, 41]. A good summary is a tradeoff between *relevance* or coherence and *diversity* or non-redundancy and is also constrained by its *length*. Thus there are three quality scores associated with a summary S with respect to a document or set of documents V: Relevance $L(S)$, Diversity $D(S)$ and Length $c(S)$. Lin and Bilmes [41] formulate the summarization problem as:

$$max\ L(S)+\lambda D(S);\ s.t.\ c(S)\leq b$$

where b is a given budget and $\lambda>0$ is a trade-off coefficient; see [41] for details. This formulation has formal similarities with standard machine learning formulation: training error *plus* regularization penalty. This kind of optimization problem is unfortunately NP-hard in general, but if the objective function is *submodular* then there is a fast scalable algorithm that returns a near-optimal i.e. $(1-1/e)$ factor approximation. Lin and Bilmes argue that the diminishing returns property make submodular functions natural in the summarization context and suggest several submodular functions for L and D suitable for the summarization problem. They also show that many existing popular approaches fall into this framework.

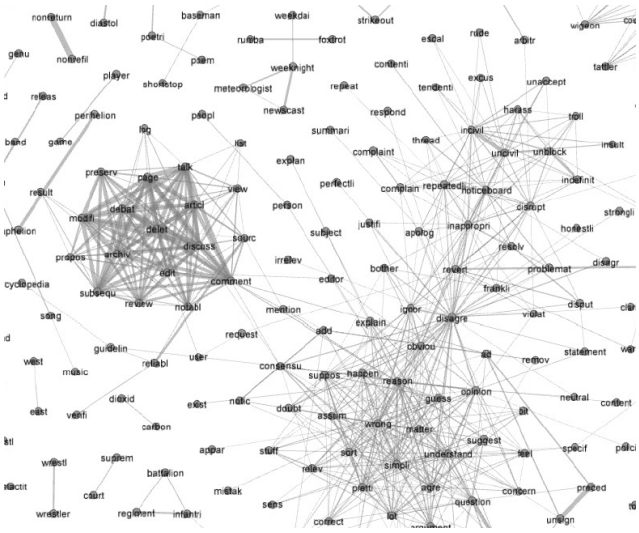

Figure 4: Word clusters based on Wikipedia

Underlying the approach of Lin and Bilmes is a notion of sentence similarity. While they employ a fast scalable method for the optimization problem, they pay relatively little attention to the similarity measure, using standard cosine similarity with *tf* and *idf*. This can be expected to suffer from the usual problems with polysemy and synonymy. Our first contribution is then to employ a knowledge-based approach to extract hidden or latent similarity

between sentences that cannot be detected using word counts alone. Sentences are given a sparse representation in terms of concepts based on a term-document matrix constructed from Wikipedia documents (Figure 4). Our second contribution is to adopt an integrative approach using multiple similarity kernels: bag of words, sparse representation of Wikipedia topics, syntactic similarity (parse tree similarity) etc. For this we use a technique we recently developed [33] based on multiple kernel learning and a classical geometrical representation of graphs; for details see [33].

6. CONCLUSIONS AND FUTURE WORK

This paper has presented some initial work and ideas developed in an ongoing project in the area of culturomics, which we believe is a research field with promising future [8] because we are faced with new bodies of evidence, a breeding ground for knowledge and new ideas and for testing new methodologies. Therefore the real challenges and opportunities lie ahead and will emerge in due course. The project will focus on the theoretical and methodological advancement of the state of the art in extracting and correlating information from large volumes of Swedish text using a combination of knowledge-based and statistical methods. One central aim of this project will be to develop methodology and applications in support of research in disciplines where text is an important primary research data source, primarily the humanities and social sciences.

The innovative core of the project will be the exploration of how to best combine knowledge-rich but sometimes resource-consuming natural language processing with statistical machine-learning and data-mining approaches. In all likelihood this will involve a good deal of interleaving. One example: Topic models based on documents as bags-of-words could be used for preliminary selection of document sets to be more deeply processed; the results of the semantic processing subsequently being fed into other machine-learning modules, back to the topic models in order to refine the selection, or for processing by semantic reasoners. In order to guide this development, visual analytics will be invaluable. If the results of processing can be represented as graphs – which are often a natural representation format for linguistic phenomena in texts or text collections – a whole battery of network analysis tools will be at the researcher's disposal. We can use them to find clusters of similar documents or entities, pinpoint those documents that were most influential to the rest, or perform any of a number of other tasks designed for network analysis.

Moreover, a goal of the project which will ensure its visibility is to replicate the IBM Watson system for Swedish with knowledge extracted from different sources such as Swedish Wikipedia, Språkbanken's contemporary and historical newspaper corpora, the classical Swedish literary works available in the Swedish Literature Bank, etc. The selection of text for the project will be restricted to readily available digital text collections and corpora. Digitization will not fall within the scope of the project. Existing text collections and corpora are already large and broad enough to ensure the viability of the methods. Furthermore, a large number of resources (e.g., Swedish FrameNet and other lexical resources) and tools are available, such as entity taggers, semantic role labelers, syntactic parsers (including converters that can transform constituent parse trees into dependency graphs) and the like. Some of the tools are at a more preliminary stage (e.g., a coreference solver [6]) while other are mature and extensively tested with various types of big data (e.g., named entities [12, 13]). Nonetheless, all resources have to be adapted and "glued" together in order to be capable of obtaining the desired goals of our culturomics project.

7. ACKNOWLEDGMENTS

The project "Towards a knowledge-based culturomics" is 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.

8. REFERENCES

[1] Acerbi, A., Lampos, V., Garnett, P., and Bentley, A. 2013. The expression of emotions in 20th century books. *PLoS ONE*. 8 (3). doi:10.1371/journal.pone.0059030. PMID 23527080.

[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th international semantic web and 2nd Asian conference on Asian semantic web conference (ISWC + ASWC)*. Busan, Korea. 722–735.

[3] Baroni, M. and Bernardini, S. (editors). 2006. *Wacky! Working papers on the Web as Corpus*. Gedit, Bologna.

[4] Berberich, K., Bedathur, S., Sozio, M., and Weikum, G. 2009. Bridging the terminology gap in web archive search. In *Proceedings of the 12th International Workshop on the Web and Databases. WebDB.* Rhode Island, USA.

[5] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia – a crystallization point for the web of data. *Journal of Web Semantics*. 154–165.

[6] Björkelund, A., Hafdell, L., and Nugues, P. 2009. Multilingual semantic role labeling. In *Proceedings of* the *15th Conference on Computational Natural Language Learning (CoNLL)*. Boulder, USA. 43–48.

[7] Blei, D. 2012. Probabilistic topic models. *Communications of the ACM*, 55, 4. doi:10.1145/2133806.2133826

[8] Bohannon, J. 2011. Google Books, Wikipedia, and the future of culturomics. *Science*. Jan 14;331(6014),135. doi: 10.1126/science.331.6014.135. 2011.

[9] Bonzanini, M., Martinez-Alvarez, M., and Roelleke, T. 2013. Extractive summarisation via sentence removal: Condensing relevant sentences into a short summary. In *Proceedings of the 36th ACM Special Interest Group on Information Retrieval (SIGIR)*. Dublin, Ireland.

[10] Borin, L., Danélls, D., Forsberg, M., Kokkinakis, D., and Toporowska Gronostaj, M. 2010. The past meets the present in Swedish FrameNet++. In *Proceedings of the 14th EURALEX International Congress*. Leeuwarden, Netherlands. 269–281.

[11] Borin, L., Forsberg, M., and Roxendal, J. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. ELRA. Istanbul, Turkey. 474–478.

[12] Borin, L., Kokkinakis, D., and Olsson, L-J. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In *Proceedings of the ACL*

Workshop: Language Technology for Cultural Heritage Data (LaTeCh). ACL. Prague, Czech Republic. 1-8.

[13] Borin, L. and Kokkinakis D. 2010. Literary onomastics and language technology. In *Literary Education and Digital Learning*, van Peer, W., Zyngier, S. and Viana, V. (eds.). Information Science Reference, Hershey - New York, 53–78. doi:10.4018/978-1-60566-932-8.

[14] Burchardt, A., Erk, K., Frank, A., Kowalski, A., and Padó, S. 2006. SALTO - A versatile multi-level annotation tool. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*. Genoa, Italy.

[15] Buyko, E., Faessler, E., Wermter, J., and Hahn, U. 2011. Syntactic simplification and semantic enrichment - trimming dependency graphs for event extraction. *Computational Intelligence* 27.4, 610–644.

[16] Chaudhry, B. 2012. Putting IBM Watson to work in healthcare. In *Analytics in Support of Health Care Transformation : Making Better Health Care Decisions with IBM Watson and Advanced Analytics*. Washington D.C., USA. <https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/Basit%20Chaudhry's%20Presentation/$file/Basit%20Chaudhry's%20Presentation.pdf>

[17] Christensen, J., Mausam, Soderland, S., and Etzioni, O. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*. Los Angeles, CA., USA. 52–60.

[18] Exner, P. and Nugues, P. 2011. Using semantic role labeling to extract events from Wikipedia. In *Proceedings of DeRiVE 2011*, Bonn.

[19] Exner, P. and Nugues, P. 2012a. Constructing large proposition databases. In *Proceedings of the Language Resources and Evaluation (LREC).* Istanbul, Turkey. 3836–3840.

[20] Exner, P. and Nugues, P. 2012b. Entity extraction: From unstructured text to DBpedia RDF triples. In *Proceedings of WoLE 2012, CEUR Workshop Proceedings*. Boston. 58–69.

[21] Exner, P. and Nugues, P. 2012c. Ontology matching: from PropBank to DBpedia. In *Proceedings of the Swedish Language Technology Conference (SLTC).* Lund, Sweden. 25–26.

[22] Fan, J., Kalyanpur, A., Gondek, D. C., and Ferrucci, D. A. 2012. Automatic knowledge extraction from documents. *IBM Journal of Research and Development*. 56,3.4, 5:1–5:10.

[23] Ferrucci, D. 2012. Introduction to 'This is Watson'. IBM *Journal of Research and Development.* 56, 3.4, 1:1–1:15.

[24] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine*, Fall 2010, 59-79.

[25] Fillmore, C., Johnson, C., and Petruck., M. 2003. Background to FrameNet. *International Journal of Lexicography*, 16, 3, 235–250.

[26] Fürstenau, H. and Lapata, M. 2009. Semi-supervised semantic role labeling. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Athens, Greece.

[27] Gildea, D. and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics*. 28,3, 245-288.

[28] Goldberg, Y. and Orwant, J. 2013. A dataset of syntactic-Ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*. Atlanta, Georgia, USA. 241–247.

[29] Halevy, A., Norvig, P., and Pereira, F. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24, 2, 8-12. doi:10.1109/MIS.2009.36.

[30] Hall, D., Jurafsky, D., and Manning, C. D. 2008. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Honolulu, USA. 363–371.

[31] Hitchcock, T. 2011. *Culturomics, Big Data, Code Breakers and the Casaubon Delusion*. <http://historyonics.blogspot.se/2011/06/culturomics-big-data-code-breakers-and.html>

[32] Hoover, Q. 2013. *Transforming Health Care Through Big Data. Strategies for Leveraging Big Data in the Health Care Industry.* Institute for Health Technology Transformation. New York, USA.

[33] Jethava, V., Martinsson, A., Bhattacharyya, C., and Dubhashi, D. 2012. The Lovasz θ function, SVMs and finding large dense subgraphs. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*. Lake Tahoe, NV; USA. 1169-1177.

[34] Jockers, M. L. 2013. *Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities)*. UIUC Press.

[35] Johansson, R., Friberg Heppin, K., and Kokkinakis, D. 2012. Semantic role labeling with the Swedish FrameNet. In *Proceedings of 8th Language Resources and Evaluation Conference (LREC)*. Istanbul, Turkey. 3697-3700.

[36] Johansson, R. and Nugues, P. 2008. Dependency-based syntactic–semantic analysis with PropBank and NomBank. In *Proceedings of Conference on Natural Language Learning (CoNLL)*. Manchester, UK. 183–187.

[37] Kaluarachchi, A., Roychoudhury, D., Varde, A.S., and Weikum, G. 2011. Sitac: discovering semantically identical temporally altering concepts in text archives. In *Proceedings of the 14th International Conference on Extending Database Technology (EDBT/ICDT)*. ACM, New York, NY, USA. 566–569.

[38] Kaluarachchi, A. C., Varde, A. S., Bedathur, S., Weikum, G., Peng, J., and Feldman, A. 2010. Incorporating terminology evolution for query translation in text retrieval with association rules. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10. ACM, New York, NY, USA. 1789–1792.

[39] Kanhabua, N. and Nørvåg, K. 2010. Exploiting timebased synonyms in searching document archives. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10. ACM, New York, NY, USA. 79–88.

[40] Leetaru, K. 2011. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*. 16, 9. Chicago University Library. Available online: <http://journals.uic.edu/ojs/index.php/fm/article/view/3663/3040#p7>

[41] Lin, H. and Bilmes, J. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT)*. Portland, Oregon.

[42] Mausam, Schmitz, M., Bart, R., Soderland, S., and Etzioni, O. 2012. Open language learning for information extraction. In *Proceeding of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Jeju, Korea. 523-534.

[43] Mazeika, A., Tylenda, T., and Weikum, G. 2011. Entity timelines: Visual analytics and named entity evolution. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, New York, NY, USA. 2585–2588.

[44] Michel, J-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. 2010. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014), 176–82. doi: 10.1126/science.1199644.

[45] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Meeting of the ACL and the 4th International Joint Conference on NLP of the AFNLP: Volume 2*. ACL, Singapore, 1003–1011.

[46] Moretti, F. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History.* R. R. Donnelley & Sons.

[47] Nguyen, T. V. T. and Moschitti, A. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the ACL*. Portland, Oregon, USA. 277–282.

[48] Nugues, P. M. 2006. *An Introduction to Language Processing with Perl and Prolog. An Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German.* Springer Verlag, Berlin Heidelberg New York.

[49] O'Reilly. 2012. *Big Data Now: 2012 Edition*. O'Reilly Media, Inc.

[50] Petersen, A. P., Tenenbaum, J., Havlin, S., and Stanley, H. E. 2012. Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports* 2, 313. doi: 10.1038/srep00313

[51] Punyakanok, V., Roth, D., and Yih, W. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*. 34, 2, 257–287.

[52] Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Keim, D. A., and Plank, F. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon. 305–310.

[53] Suchanek, F. M., Kasneci, G., and Weikum, G. 2007.Yago - A core of semantic knowledge. In *Proceedings of the 16th international World Wide Web conference (WWW07)*. Alberta, Canada. 697–706.

[54] Zimmer, B. 2011. Twitterology: A new science? *The New York Times - Sunday Review.* <http://www.nytimes.com/2011/10/30/opinion/sunday/twitterology-a-new-science.html?_r=0>