JPMTR 022 | 1311
UDC xxxxx

Original scientific paper
Received: 2013-06-28
Accepted: 2013-11-07

Software Newsroom - an approach to automation of news search and editing

Juhani Huovelin[1], *Oskar Gross*[2], *Otto Solin*[1], *Krister Lindén*[3], *Sami Maisala*[1], *Tero Oittinen*[1],
Hannu Toivonen[2], *Jyrki Niemi*[3], *Miikka Silfverberg*[3]

[1] Division of Geophysics and Astronomy
Department of Physics, University of Helsinki
FIN-00560 Helsinki, Finland

E-mails: juhani.huovelin@helsinki.fi
otto.solin@helsinki.fi
sami.maisala@helsinki.fi
tero.oittinen@helsinki.fi

[2] Department of Computer Science and HIIT
University of Helsinki
FIN-00014 Helsinki, Finland

E-mails: oskar.gross@cs.helsinki.fi
hannu.toivonen@cs.helsinki.fi

[3] Department of Modern Languages
University of Helsinki
FIN-00014 Helsinki, Finland

E-mails: krister.linden@helsinki.fi
jyrki.niemi@helsinki.fi
miikka.silfverberg@helsinki.fi

Abstract

We have developed tools and applied methods for automated identification of potential news from textual data for an automated news search system called Software Newsroom. The purpose of the tools is to analyze data collected from the internet and to identify information that has a high probability of containing new information. The identified information is summarized in order to help understanding the semantic contents of the data, and to assist the news editing process.

It has been demonstrated that words with a certain set of syntactic and semantic properties are effective when building topic models for English. We demonstrate that words with the same properties in Finnish are useful as well. Extracting such words requires knowledge about the special characteristics of the Finnish language, which are taken into account in our analysis. Two different methodological approaches have been applied for the news search. One of the methods is based on topic analysis and it applies Multinomial Principal Component Analysis (MPCA) for topic model creation and data profiling. The second method is based on word association analysis and applies the log-likelihood ratio (LLR). For the topic mining, we have created English and Finnish language corpora from Wikipedia and Finnish corpora from several Finnish news archives and we have used bag-of-words presentations of these corpora as training data for the topic model. We have performed topic analysis experiments with both the training data itself and with arbitrary text parsed from internet sources. The results suggest that the effectiveness of news search strongly depends on the quality of the training data and its linguistic analysis.

In the association analysis, we use a combined methodology for detecting novel word associations in the text. For detecting novel associations we use the background corpus from which we extract common word associations. In parallel, we collect the statistics of word co-occurrences from the documents of interest and search for associations with larger likelyhood in these documents than in the background. We have demonstrated the applicability of these methods for Software Newsroom. The results indicate that the background-foreground model has significant potential in news search. The experiments also indicate great promise in employing background-foreground word associations for other applications.

A combined application of the two methods is planned as well as the application of the methods on social media using a pre-translator of social media language.

Keywords: social media, data mining, topic analysis, machine learning, word associations, linguistic analysis

1. Introduction

The vast amount of open data in the internet provides a yet ineffectively exploited source of potential news. Social media and blogs have become an increasingly useful and important source of information for news agencies and media houses. In addition to the news collected, edited and reported by traditional means, i.e., by news agencies, the information in a news-room consists of different types of user inputs. In the social media there is a large amount of user comments and reactions triggered by news stories. Also, fresh article manuscripts and other types of material can be produced by basically anyone by submitting the information to the internet. As a means of collecting news, this material is already in use by commercial media companies, especially in a hyperlocal media context (e.g., newspapers that discuss local issues).

While this editorial strategy is considerably more advanced than the way news were produced a decade ago, the work still includes manual work that could be automated and the use of open data available in the internet is usually very inefficient. It also does not make much sense to engage humans for browsing internet data, a job that can be done much more efficiently and tirelessly by a machine.

Thus, intelligent computer algorithms that monitor internet data and hunt for anomalies and changes are becoming an increasingly exploited means of news and trend detection. Other applications for the same methodologies are public opinion analysis and forecasting the results of elections. Examples of even more advanced intelligence in prediction would be calls to events, which can be predecessors of demonstrations or even an uprising, and indication of a meeting between high level politicians based on their plans to travel to the same place at the same time.

The same methods, when combined with fusion of heterogeneous data, can help improving the quality and widening the scope of news by the enrichment of existing news material with relevant background information and other associated material (e.g., history, pictures, digital video material). In principle, using the same methodology it is also possible to follow the discussions raised by published news articles and thus automatically collecting feedback from the audience.

Examples of internet services developed for the above purposes are Esmerk Oasis (Comintelli, 2013) and Meltwater Buzz (Meltwater, 2013). Esmerk Oasis is a web-based market intelligence solution. Its services include customized global business information with the possibility of importing complementary information from other sources as well as sharing and distribution of information across the client organization. Meltwater Buzz is a social media monitoring tool that has capabilities for tracking and analyzing user-generated content on the web. Google has also developed several services that perform similar tasks.

Considering the purpose and goal of an automated news search and analysis process, a baseline approach to analyzing text material and creating a short descripttion of its contents is to simulate the traditional process of news production. The analysis of the material should tell you *what, who, where, and when*? Methodologically, the most challenging task is to find a systematic way of defining the answer to the question *what*, since it includes the need to recognize and unambiguously describe an unlimited range of *topics,* not just individual words. A topic is usually defined as "a set of news stories that are strongly related by some seminal real-world event", and an event is defined as "something (non-trivial) happening in a certain place at a certain time" (Allan, 2002). As an example, the recent meteorite impact in Chelyabinsk was the event that triggered the asteroid impact, natural catastrophes, and doomsday topic. All stories that discuss the observations, consequences, witnesses, probabilities and frequency of such events etc., are part of the topic.

The answers to the other questions*, who, where* and *when,* can be traced by searching named entities and various time tags and information. In practical application to, e.g., social media, however, the latter questions may also pose a significant challenge for an automated approach, since social media language does not obey common rules.

The quality of the language is often very poor, since it may include many local and universal slang words, acronyms and idioms that are known by only a limited local community, and also numerous typing errors.

Blogs are considerably less difficult in this respect, since most of the text in them is in fairly well written standard language.

Methods for event and trend detection and analysis in large textual data include *static and dynamic component models* which are well suited for news search and detection in the internet. Static models are simpler to use and give results that are easier to interpret. A potential disadvantage is that newly emergent trends may remain undetected if the training data for the model is not sufficiently extensive, leading to the model being not generic enough. A dynamic model, on the other hand, is updated continuously in order to keep up with possible emergent topics. Its usage, however, is not as straightforward as that of static models since the emergent trends may be described in terms of dynamic components whose semantics is not yet well understood.

An example of a static component model is Principal Component Analysis (PCA). PCA was invented in 1901 by Pearson (1901). PCA can be performed by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix. The singular value decomposition of the word count matrix is also cal-led Latent Semantic Indexing (LSI) (Berry, Dumais and O'Brien, 1994; Hofmann, 1999).

We have developed algorithms for automated analysis of text in, e.g., social media, blogs and news data with the aim of identifying "hot" topics that are potential news. We here present the methods and show results of their application using real data.

2. Method

2.1 Combining methods

We apply two different approaches that are combined in order to achieve a clearer recognition of potential news in an arbitrary text under analysis. The first method is topic mining including advanced linguistic analysis for named entity recognition. The topic model is based on Multinomial Principal Component Analysis, MPCA (Kimura, Saito and Uera, 2005; Buntine and Jakulin, 2006). While topics are considered to be different kinds of objects than named entities (e.g., Newman et al., 2006), they can be combined in the creation of a probabilistic topic model. The second approach, association analysis, takes into account the word co-occurrences in the document and uses statistics to look for novel word associations in a set of documents. These associations are used for Software Newsroom applications, such as diverging (association) word clouds and automatic summary generation. The background model calculation uses a method based on the log-likelihood ratio (LLR) (Dunning, 1993). This is described in more detail by Toivonen et al. (2012). By extending the ideas in the latter approach, we propose a method for detecting novel word associations.

2.2 Topic mining

For generating static component models from textual data, we use the statistical generative model called *Multinomial principal component analysis* (MPCA) (Buntine and Jakulin, 2006). MPCA is used to model the data in order to obtain a comprehensive understanding of the contents of the data sources in the form of semantically meaningful components or topics.

In our application, the topic model includes four categories of common words (nouns, verbs, adjectives, adverbs) where the nouns are not named entities, and four categories of named entities (persons, places, organisations, miscellaneous), where the miscellaneous category includes all named entities that do not belong to the other three categories. It has been shown that these eight categories are effective for building topic models for English (e.g., Newman et al., 2006). An important aspect of our research is to verify that the linguistic categories can be identified in a language-independent way. We demonstrate this by extracting the eight categories of words from text in Finnish - a language completely unrelated to English.

Let $\mathbf{D}$ be a $d \times N$ matrix representing the training data (documents) as a "bag-of-words", $\mathbf{M}$ a $d \times K$ matrix of documents represented in terms of topics, and $\boldsymbol{\Omega}$ a $K \times N$ matrix of topics represented in terms of words, where d is the number of documents in the training corpus, N the size of the vocabulary and K the number of topics ($K << N$). We extract from the training corpus two types of features: Part-of-speech tags (nouns, adjectives, verbs, adverbs) and Named Entities (locations, persons, organizations, miscellaneous). Thus, in our case, the vocabulary words are treated as eight multinomials. The aim is to represent the documents in terms of matrices $\mathbf{M}$ and $\boldsymbol{\Omega}$ as (Equation 1):

$$\mathbf{D} \approx \mathbf{M} \times \boldsymbol{\Omega}. \qquad [1]$$

In other words, the data is transformed into a lower dimensional space, where documents are represented in terms of topics. The topics are then represented in terms of words. The matrices $\mathbf{M}$ and $\boldsymbol{\Omega}$ give the probabilities of topics given a document and words given a topic, respectively.

The process for generating the model with MPCA is as follows (Buntine and Jakulin, 2004).

1. A total of N words are partitioned into K partitions $\mathbf{c} = c_1, c_2, \ldots, c_K$ where $\sum_{k=1}^{K} c_k = N$. N is the size of the vocabulary and K the number of topics. The partitioning is done using a latent proportion vector $\mathbf{m} = (m_1, m_2, \ldots, m_K)$. The vector $\mathbf{m}$ for each document forms the d rows in matrix $\mathbf{M}$.

2. Words are sampled from these partitions according to the multinomial for each topic producing a bag-of-words representation $\mathbf{w}_{k,\cdot} = (w_{k,1}, w_{k,2}, \ldots, w_{k,N})$ for each partition k.

3. Partitions are combined additively to produce the final vocabulary $\mathbf{r} = (r_1, r_2, \ldots, r_N)$ by totaling the corresponding counts in each partition,

   $r_n = \sum_{k=1}^{K} w_{k,n}$

The above process is described by the following probability model (Equations 2),

$$\begin{aligned} &\mathbf{m} \sim \mathit{Dirichlet}(a) \\ &\mathbf{c} \sim \mathit{Multinomial}(\mathbf{m}, N) \qquad [2] \\ &\mathbf{w}_{k,\cdot} \sim \mathit{Multinomial}(\boldsymbol{\Omega}_{k,\cdot}, c_k) \quad \text{for } k = 1,\ldots,K \end{aligned}$$

Its estimation is done through a Gibbs sampler (Buntine and Jakulin, 2006). In Gibbs sampling, each unobserved variable in the problem is resampled in turn according to its conditional distribution. Its posterior distribution conditioned on all other variables is computed, and then a new value for the variable using the posterior is sampled. In each cycle of the Gibbs algorithm the last $\mathbf{c}$ for each document is retrieved from storage and then, using a Dirichlet prior for rows of $\boldsymbol{\Omega}$, the latent component variables $\mathbf{m}$ and $\mathbf{w}$ are sampled. The latent variables are $\mathbf{m}$ and $\mathbf{w}$, whereas $\mathbf{c}$ is derived. As a result we get estimates of the matrices $\mathbf{M}$ and $\boldsymbol{\Omega}$ in

Equation 1. In the context of an MPCA model, these are estimates of the distribution of documents over topics and topics over words respectively.

There are different ways of estimating topic strengths in a single document given the model created by MPCA. The method applied here is cosine similarity between document vector *d* and topic ω_k as

$$\mathrm{sim}(\mathrm{d},\omega_\mathrm{k}) = \frac{\mathrm{d}\cdot\omega_\mathrm{k}}{||\mathrm{d}||||\omega_\mathrm{k}||}. \qquad [3]$$

A topic model including a desired number of yet unnamed topics is first created by the above method (Equations 1 and 2) using a bag-of-words presentation of the training corpus. This is then ready for application in topic analysis of arbitrary text. The topic analysis includes automated simultaneous identification of the topic, person, place, organization, and event in an arbitrary blog article, a discussion thread in social media or an RSS feed, etc. This is done by statistical comparison, or projection (Equation 3), of the new text against the topic model. By tracking the history of the frequency of occurrence of similar stories (which belong to the same topic, i.e., resemble each other), the software can identify the trend of a topic. A statistically significant deviation from the trend in a short time period gives a hint that the source texts that caused this deviation may include a news candidate. In the present analysis, we use Gaussian statistics and the criteria for significant deviation is 3σ. This applies to generic topics, but for words and events that are generally interesting from a newsroom perspective, such as VIPs, accidents, crimes, and natural disasters, all occurrences are tagged as being potential news topics. The method is, on a general level, similar to the approach of Newman et al. (2006) but it includes advanced features developed for practical applicability in a newsroom environment.

2.3 Linguistic analysis for named-entity recognition

2.3.1 English vs. Finnish words and named entities

When adapting topic identification from one language to another, it is necessary to be aware of what units of the language have been chosen and how similar units can be identified in another language. All language analysis methods do not produce the same output granularity. In the following, we outline the units that have been found effective in English and how corresponding units can be identified in Finnish to highlight some of the essentials that need to be considered when choosing linguistic analysis software to adapt to another language.

The most striking difference between Finnish and English is the number of inflected forms in Finnish. There are roughly 2000 forms for each noun, 6000 for each adjective and 12000 for each verb. The characteristics of these forms and their usage in Finnish has been extensively documented in an online Finnish grammar, "Iso suomen kielioppi" (Hakulinen et al., 2004). It is not possible to only chop off word endings, because changes also take place in the stem when inflectional morphemes are added, e.g., "nojatuoli" [armchair], "nojatuoleja" [armchairs], "nojatuoleissa" [in the armchairs], "nojatuoleissani" [in my armchairs], "nojatuoleissanikin" [also in my armchairs]. In practice, Finnish words can represent expressions that in English are rendered as a phrase, so Finnish needs a morphological analyzer to separate the base form from the endings. As a bonus to the morphological processing, many of the inflectional morphemes that are separated from the base form correspond to stop-words in English.

In addition to gluing inflectional morphemes onto the words, Finnish also has the orthographic convention of writing newly formed compound words without separating spaces, i.e., "nappanahkanojatuoli" [calf-skin armchair]. The English word *armchair* can be seen as a compound as well, but typically a modern *armchair* is not perceived only as a *chair* with *arm*rests, but as something slightly more comfortable, so the *armchair* has a lexicalized meaning of its own. This means that, for newly coined non-lexicalized compounds, it is essential that the morphological analysis separates the non-lexicalized parts in Finnish; otherwise the compositional meaning is lost. Long newly formed compounds also lack predictive power since they are rare by definition whereas the compound parts may give essential clues to the topic of the narrative. It should be noted that a Finnish writer could also choose to write "nappanahkainen nojatuoli" [calf-skin armchair], and with the increased influence of English, this convention is perceived as more readable.

The structure of named-entities, i.e., places, organizations, persons and other names, follows the conventions mentioned for regular words. In particular, place names tend to be written in one or two words at most because they are of older origin. Person names have a similar structure as in English with given name and surname. However, long organization names tend to be formulated as multi-word expressions following newer writing tendencies.

2.3.2 Named-entity recognition in Finnish

For named-entity recognition in many languages it is possible to do string matching directly on the surface forms in written text. In Finnish, we need more in-depth morphological processing to deal with the inflections and the compound words. For out-of-vocabulary words, we also need guessers. To cope with morphological ambiguity, we need a tagger before we can apply named-entity recognition.

Language technological applications for agglutinating languages such as Finnish, benefit greatly from high coverage morphological analyzers providing word forms with their morphological analyses, e.g.,

> "nojatuole+i+ssa+ni+kin : nojatuoli *Noun Plural 'In' 'My' 'Also'*" [also in my armchairs].

However, morphological analysis makes applications dependent on the coverage of the morphological analyzer. Building a high coverage morphological analyzer (with an accuracy of over 95%) is a substantial task and, even with a high-coverage analyzer, domain-specific vocabulary presents a challenge. Therefore, accurate methods for dealing with out-of-vocabulary words are needed.

With the Helsinki Finite-State Transducer (HFST) tools (Lindén et al., 2011), it is possible to use an existing morphological analyzer for constructing a morphological guesser based on word suffixes. Suffix based guessing is sufficient for many agglutinating languages such as Finnish (Lindén and Pirinen, 2009), where most inflection and derivation is marked using suffixes. Even if a word is not recognized by the morphological analyzer, the analyzer is likely to recognize some words which inflect similarly as the unknown word. These can be used for guessing the inflection of the unknown word.

Guessing of an unknown word such as "twiitin" (the genitive form of "twiitti", tweet, in Finnish) is based on finding recognized word forms like "sviitin" (genitive form of "sviitti" hotel suite in Finnish), that have long suffixes such as "-iitin", which match the suffixes of the unrecognized word. The longer the common suffix, the likelier it is that the unrecognized word has the same inflection as the known word. The guesser will output morphological analyses for "twiitin" in order of likelyhood.

A morphological reading is not always unique without context, e.g., "alusta" can be an inflected form of "alku" [beginning], "alunen" [plate], "alustaa" [found] or "alus" [ship]. To choose between the readings in context it is possible to use, e.g., an hidden Markov model (HMM) which is essentially a weighted finite-state model. Finite-state transducers and automata can more generally be used for expressing linguistically relevant phenomena for tagging and parsing as regular string sets, demonstrated by parsing systems like Constraint Grammar (Karlsson, 1990) which utilizes finite-state constraints. Weighted machines offer the added benefit of expressing phenomena as fuzzy sets in a compact way.

Using tagged input, a named entity recognizer (NER) for Finnish marks names in a text, typically with information on the type of the name (Nadeau and Sekine, 2007). Major types of names include persons, locations, organizations and events. NER tools often also recognize temporal and numeric expressions. NER tools typically use gazetteers, lists of known names, to ensure that high-frequency names are recognized with the correct type. For Finnish, the gazetteer is included in the morphological analyzer because names inflect. In addition, names and their types can be recognized based on internal evidence, i.e., the structure of the name itself (e.g., ACME Inc., where *Inc.* indicates that ACME denotes a company), or based on external evidence, i.e., the context of the name (e.g., *works for* ACME; ACME *hired a new CEO*) (MacDonald, 1996).

2.4 Association analysis

2.4.1 Extracting word associations

One of the goals of the Software Newsroom is to give an overview of popular topics discussed in the internet communities. This gives journalists an opportunity to react to these topics on a short notice. In the Software Newsroom, word association analysis is used for detecting novelty in the contents of a given set of documents. For instance, consider a web forum where people discuss about different topics, e.g., fashion, technology, politics, economics, computer games, etc. As an example, consider that a new smartphone *SoftSmart* has a feature which automatically disables GPS when you are indoors. It turns out that it has a bug, and in some very specific cases (e.g., for instance when you are on the top floor of a building) it starts to drain your battery because the signal strength is varying. It is reasonable to believe that many *SoftSmart* users will go to web forums and start discussing about the problems. Even more, it might turn out that there is an easy fix available and this is posted somewhere to the forum. The problem is, that there are thousands of similar problems being discussed all over the world, so it is not feasible for a technology journalist to monitor all the forums.

If we could automatically detect this as a trendy topic, then this information would be invaluable for a technology journalist, as she/he could then learn more about this and write a news story. From the language analysis point of view, the text written by people in web forums and other web communities introduce problems - the text contains slang, typing errors, words from different languages, etc. These aspects add another goal for the association analysis - our goal is to develop a method which is not fixed to any specific vocabulary. Our idea is to analyze the associations between words and to look for such associations which are novel with regards to other documents.

Considering the *SoftSmart* example, there are words which co-occur in sentences but the association between them is most probably very common, such as *SoftSmart - battery, battery - drain, SoftSmart - GPS,* etc. For the *SoftSmart* case, the words for which the association is rather specific could be *battery - floor, floor - drain, battery - top, Softsmart - floor* and so on. In association analysis, our goal is to automatically detect the latter ones. Note, that the association itself might be surprising, though it is between very common words, like 'battery' and 'floor'.

Finding associations between concepts which can be represented as sets of items is a very much studied area which originates from the idea of finding correlations in market basket data (Agrawal et al., 1993). The bag-of-words model of representing documents as sets of unordered words is a common concept in information retrieval (Harris, 1954; Salton, 1993). Often, the bag-of-words model is used together with the tf-idf measure that measures word specificity with respect to the document corpus (Salton, 1993).

Analysing word associations in document is not a new idea. There are various word association measures available - the log-likelihood ratio test (Dunning, 1993), the chi-squared test, Latent Semantic Indexing (Dumais et al., 1988), pointwise mutual information (Church and Hanks, 1990), Latent Dirichlet Allocation (Blei et al., 2003), etc. There is also a method for pairs, which is inspired by tf-idf, called tpf-idf-tpu which is a combination of using term pair frequency, its inverse document frequency and the term pair uncorrelation for determining the specific pairs of a document (Hynönen et al., 2012).

In this paper, we present a method for analyzing and representing documents on the word association level. We use the log-likelihood ratio as the basis for our method. As mentioned before, finding associations between documents is a very common concept and the main goal for all the methods is to discover statistically strong associations between words. In some instances we are interested in such associations that are specific to a certain set of document. For instance, consider a set of documents about the singer Freddie Mercury. Imagine, that we create pairs of all the words which co-occur in the same sentence and the weight is determined by their co-occurrence statistics (e.g., weighted by the log-likelihood ratio test). Now, if we order the pairs decreasingly by association strength, we will most probably obtain pairs such as: 'freddie'-'singer', 'freddie'-'aids', 'freddie'-'bohemian', 'aids'-death', 'aids'-sick' etc. The point here is that some of the associations are important and relevant to the document set (e.g., the first two). On the other hand, the last two associations between words are very common. And this is defines our goal - we are looking for word associations which are specific to a certain set of documents *and* at the same time are uncommon with respect to other documents.

In the following, we introduce methods for extracting word associations that are specific to a set of documents. For this we define two concepts: *background associations*, which are the common associations between words and *foreground associations*, where the weight is higher for associations that are novel with respect to the background associations.

After we have given an overview of the core methods, we will present applications of these models in the Software Newsroom. First we will look at the possible representations of foreground associations and discuss the possible usefulness of explicit graph representations. Then we will provide an idea of diverging word clouds which illustrate word associations rather than frequencies. Finally, we propose a simple, yet intuitive way of generating summaries of a set of documents by using foreground associations.

2.4.2 Background associations

Background associations represent common-sense associations between terms, where the weight depends on the strength of the association. For example, the connection between the words 'car' and 'tire' should be stronger than the connection between 'car' and 'propeller'. In our methodology, these associations are extracted from a corpus of documents, motivated by the observation that co-occurrence of terms tends to imply some semantic relation between them (slightly misleadingly often called semantic similarity). Background associations are calculated by identifying words which co-occur in the same sentence. The strength between the words is calculated using the log-likelihood method (Dunning, 1993; Toivonen et al., 2012). The latter paper describes how the word associations are calculated and also demonstrates the relationship between such associations and relations in WordNet (Miller, 1995).

2.4.3 Foreground associations

In contrast to the common associations in the background, *foreground associations* represent novel associations of a (small) set *F* of documents called the foreground documents.

However, the background associations do have a central role here: they tell us what is known, so that we can infer what is novel in any given document. The weighting scheme in the foreground also uses the log-likelihood ratio test. However, now we use the background to obtain the expected number of co-occurrences and to see how much the observed number of co-occurrences in the foreground documents F deviates from it. The result of this test gives higher weights to those term pairs that are more frequent in the foreground *F* than they are in the background, i.e., especially those pairs which have a small likelihood of occurring together in the background.

In our implementation of this idea, the foreground weights are based on the log-likelihood ratio where the alternative model is based on the foreground documents *F* and the null model on the background corpus *C*.

Let parameters p_{ij}^{null} be the maximum likelihood parameters for the corpus C, i.e., (Equation 4):

$$\begin{aligned} p_{11}{}^{\text{null}} &= p(x \wedge y;\ C) \\ p_{21}{}^{\text{null}} &= p(x \wedge \neg y;\ C) \\ p_{12}{}^{\text{null}} &= p(\neg x \wedge y;\ C) \\ p_{22}{}^{\text{null}} &= p(\neg x \wedge \neg y;\ C) \end{aligned} \quad [4]$$

where x and y denote the events that "word x (respectively y) occurs in a randomly chosen sentence (of the given corpus)". For the background associations, these parameters are used as the alternative model, and here they are used as the null model. Set the alternative model parameters p_{ij} in turn to be the maximum likelyhood parameters for the document set F (Equations 5),

$$\begin{aligned} p_{11} &= p(x \wedge y;\ F) \\ p_{21} &= p(x \wedge \neg y;\ F) \\ p_{12} &= p(\neg x \wedge y;\ F) \\ p_{22} &= p(\neg x \wedge \neg y;\ F) \end{aligned} \quad [5]$$

The log-likelihood ratio (LLR) for the foreground associations is then computed according to Equation 6.

$$\text{LLR}(x, y) = -2 \sum_{i=1}^{2} \sum_{j=1}^{2} k_{ij} \log(p_{ij}{}^{\text{null}} / p_{ij}) \quad [6]$$

The foreground association weights are assigned by this LLR function. Using this function, we give higher weight to such associations which are more likely to appear in the foreground and less likely in the back-ground. Note, that the log-likelihood ratio could be also negative. In this case the word association is weaker in the foreground than in the background. In our work we omit associations with negative weights.

2.4.4 Applications

In the following, we present applications in the Software Newsroom that employ the background/foreground associations method. In the first application we describe, the associations in the set of documents are represented as an explicit graph. In the remainder of the subsection we will demonstrate two different Software Newsroom applications - diverging word cloud generation and document summarization. For a single document experiment we will use the English Wikipedia as the background corpus and a story from BBC: "Google tests balloons to beam internet from near space" (Kelion, 2013) as the foreground document we are interested in.

The simplest way of representing the information is by showing the top-k (where k is an integer) word pairs of the news story. In order to show the differences between a standard co-occurrence calculation and our foreground method in we have, in Table 1, presented the top-5 pairs of the Google news story. For comparison, the left column lists the most strongly associated word pairs as measured using standard methods, while the right column lists the top-5 pairs obtained by the foreground method.

The pairs suggest that the foreground method is able to grasp the main associations of the news story better than the classical co-occurrence measures. By this we mean that the associations of the foreground contain more relevant associations, such as 'superpressure' and 'balloons' or 'google' and 'balloons'. Representing associations as a simple list makes them individually easy to understand, but does not give a picture of the network of connections. On the other hand, a graphical representation (Figure 1) of this network may be difficult, especially for novice users. On the other hand, when a user is familiar with such data representation it gives a quick and general view of the data. In our work, the explicit graph is not a favored method for illustrating or representing information. We put more emphasis on designing methods that employ the foreground graph.

Table 1: The top-5 pairs for the BBC news story "Google tests balloons to beam internet from near space". The left column shows pairs calculated using the standard co-occurrence calculation method (log-likelihood ratio); the right column shows the top-5 pairs obtained using the foreground association method

Long-likelihood ratio	**Foreground method**
plastic - made	superpressure - balloons
months - airborne	launched - new_zealand
suggested - atmosphere	google - balloons
special - fitted	suggested - atmosphere
force - air	force - air

We now propose a new type of word clouds, *diverging (association) clouds,* that aim at helping users to explore the novel associative knowledge emerging from textual documents. Given a search term, the diverging cloud of a document highlights those words that have a special association with the search term. As a motivating application, consider word clouds as summaries of news stories. If the user has a special interest, say 'iPhone', we would first of all like the word clouds to be focused or conditioned on this search term, i.e., only show terms to which 'iPhone' is associated in the news story. Secondly, we would like to see only novel information about the iPhone, not the obvious ones such as 'Apple' and 'mobile'. The diverging clouds aim to do exactly this, directly based on the foreground associations of a news story as a representation of potentially new semantic associations. For a sample of diverging (association) clouds, see Figure 4 in the Results section.

In news, it is very common that the information on a certain event comes in over time. This is even more so for news published on the web or discussed in internet forums. For instance, considering an incident (e.g., the Boston Marathon bombing) which has a large impact and is related to many people, information and updates concerning the event are usually published frequently on news websites. For each news story update, most parts remain the same, some of the information changes, something is added, and something is removed. To get an overall picture of the event, one should go through many articles and collect the new bits of information from each of them while at the same time most of the information is redundant.

Automatic document summarization is a method for overcoming this problem. The main goal of document summarization is to represent the information in a set of documents in a short and possibly non-redundant manner.

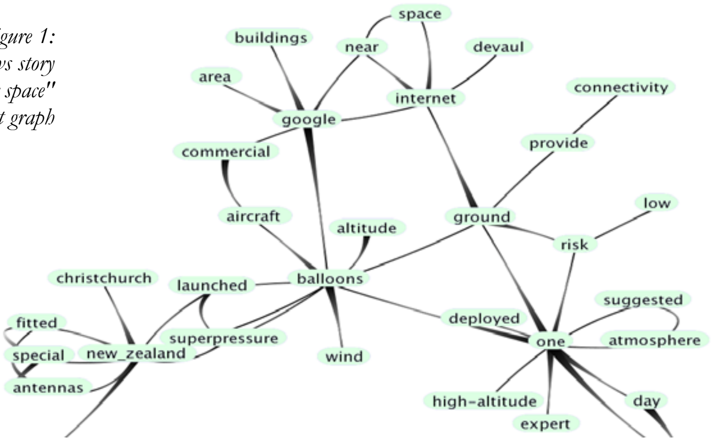

Figure 1:
A subset of the foreground associations for the news story "Google tests balloons to beam internet from near space" represented as an explicit graph

Two different approaches have been used in document summarization - generating text using the documents as reference (Hori et al., 2003; Knight et al., 2002) and selecting a representative set of sentences from the documents, e.g., by using Support Vector Machines (Yeh et al., 2005), Hidden Markov Models (Conroy and O'leary, 2001), and Conditional Random Fields (Sheh et al., 2007).

In this paper, we sketch a tentative method which uses sentence extraction for document summarization using foreground associations as reference. In the future, we plan to enhance this method and also combine it with text generation into a hybrid method. Sentence selection is based on the principle that our goal is to cover as much of the foreground associations as possible with the minimum number of sentences. The intuitive idea is that the foreground associations describe the most relevant aspects of the document set. Now, if these associations are covered by a certain set of documents, it is reasonable to assume that we have also captured the essentials of these documents.

A greedy algorithm for selecting the sentences is the following:

1. Select the strongest association *e* from the foreground associations.
2. Let *S(e)* be the set of all sentences which contain both words of *e*.
3. For each sentence *s* in *S(e)* calculate the score of the sentence as the sum of the foreground association weights for all the word pairs found in the sentence.
4. Output the highest scoring sentence *s**.
5. Remove *e* and all other pairs which appear in *s** from the foreground associations.
6. If the size of the summary is not sufficient and the foreground is not empty, go to (1).

The design of the algorithm follows the following two principles. First, the highest weighted pair should contain the most important information. This is the reason why we start looking for the sentences where the words of the highest association in the foreground appear. The second principle is that we do not want to include information which is already included. This is the reason for the step (5) above. If we are interested in penalizing sentences which contain pairs which are already covered, then in step (5) it is possible to change the values of the association weights into negative constants. For our experiments with summary generation on the topic of Google balloons, see the Results section 3.2.2.

3. Results

3.1 Topic analysis

We have applied topic mining to both the English and the Finnish languages. Our first experiment was to construct a model using 2 million articles from the English Wikipedia (at the time this comprised 13 % of the entire English Wikipedia).

The vocabulary is created by extracting eight features from the raw text divided into subdocuments. These features are based on part-of-speech (POS) classification (extracting and lemmatizing nouns, verbs, adjectives and adverbs) and named entity recognition (NER) (tagging words and groups of words as persons, locations, organizations and miscellaneous). For POS tagging we use FreeLing (Padró et al., 2010) and for NER tagging the Illinois Named Entity Tagger (Ratinov and Roth, 2009). The documents and features are forwarded to the model trainer MPCA as a bag-of-words presentation. The MPCA produces the K (in these examples K=50) strongest topics for the user to name. This name is not used for the projection of text against the model (Equation 3), but it associates a numbered topic to a semantically meaningful context, which is essential for humans who exploit the method. Tables 2 and 3 present one of the fifty topics generated. The topic has intuitively been given the name "Space missions". Documents/texts under analysis are projected against the created model in order to find which topics the text is most strongly related to. Feature extraction and bag-of-words presentation are applied to the single document (as is done for the entire corpus in the model creation) before applying Equation 3 to the projection.

Table 2:
The fourteen strongest Named Entity tags for the topic "Space missions". The un-normalized weighting factor corresponds to the incidence of the Named Entity in the particular topic. LOC stands for location, MISC for miscellaneous, ORG for organization, and PER for person. The weight is given at the left side of each word

Weight	LOC	Weight	MISC	Weight	ORG	Weight	PER
14.25	Russia	34.11	Russian	5.71	NASA	1.51	Venus
6.31	Moscow	11.34	Soviet	5.27	Sun	0.77	Ivan
5.01	Earth	6.98	Ukrainian	1.84	Apollo	0.59	Pluto
4.85	Ukraine	2.09	Estonian	1.13	Mars	0.53	Mars
4.28	Soviet Union	1.98	Georgian	1.05	Moon	0.43	Galileo
1.78	Kiev	1.87	Russians	0.96	Saturn	0.42	Mercury
1.70	Estonia	1.56	Latvian	0.75	NGC	0.38	Moon
1.66	Mars	1.05	Soyuz	0.64	Nikon	0.38	Vladimir
1.56	Jupiter	1.03	Belarusian	0.61	ISS	0.35	Ptolemy
1.52	USSR	0.74	Titan	0.57	GPS	0.34	Kepler
1.35	Georgia	0.62	Martian	0.53	ESA	0.31	Lenin
1.30	Belarus	0.62	Gregorian	0.52	Gemini	0.30	Boris
1.31	Latvia	0.54	Chechen	0.50	Canon	0.28	Koenig
1.17	Saint Petersburg	0.50	Earth	0.44	AU	0.28	Star

Table 4 shows an example based on the BBC article entitled "Storm Sandy: Eastern US gets back on its feet" (31 October 2012). Table 4 presents the five strongest topics given by the model for this news article. The numbers in front of the topics are normalized statistical weights of each topic. Table 5 presents the Named Entities given by the NER tagger for this news article.

Table 3: The strongest Part of Speech (POS) tags for the topic "Space missions". JJ stands for adjective, NN for noun, RB for adverb and VB for verb. The weight is given at the left side of each word

Weight	JJ	Weight	NN	Weight	RB	Weight	VB
2.63	solar	2.06	star	15.16	man	2.85	see
1.90	light	1.89	space	2.50	approximately	1.93	take
1.71	lunar	1.37	system	2.34	away	1.52	discover
1.41	html	1.34	planet	1.98	z_times	1.43	show
1.40	russian	1.11	object	1.66	close	1.36	give
1.08	red	1.06	camera	1.56	actually	1.32	move
1.06	astronomical	0.96	light	1.47	relatively	1.30	name
1.05	bright	0.93	satellite	1.46	slightly	1.28	find
1.04	scientific	0.87	crater	1.44	probably	1.26	appear
1.04	black	0.86	day	1.22	roughly	1.13	observe
1.03	dark	0.84	mission	1.20	sometimes	1.11	launch
0.99	similar	0.81	orbit	1.19	currently	1.01	call
0.97	visible	0.80	distance	1.16	long	0.86	refer
0.90	optical	0.75	lens	1.16	directly	0.77	base

Table 4: The strongest topics of the BBC, 31 October 2012 article "Storm Sandy: Eastern US gets back on its feet". The normalization is such that the total sum of the weights of all words in the material is 1

	Weight	Topic Name
1	0.0512	US politics
2	0.0511	Sci-fi and technology
3	0.0391	US traffic and information networks
4	0.0384	Latin America
5	0.0342	Physics

Table 5: The Named Entities for the BBC news article "Storm Sandy: Eastern US gets back on its feet" (31 October 2012)

Freq.	Type	Entity	Freq.	Type	Entity
1	MISC	Democratic	1	PER	Andrew Cuomo
1	MISC	Earth A	1	PER	Barack Obama
1	MISC	Jersey Shore	2	PER	Chris Christie
1	MISC	Nasdaq	2	PER	Christie
3	MISC	Republican	1	PER	Donna
1	LOC	Atlantic City	1	PER	Joseph Lhota
1	LOC	Canada	1	PER	Michael Bloomberg
1	LOC	Caribbean	1	PER	Mitt Romney
1	LOC	Easton	1	PER	Mt Washington
1	LOC	Haiti	2	PER	Obama
1	LOC	Hudson River	1	PER	Paul Adams
1	LOC	JFK	1	PER	Romney
4	LOC	Manhattan	6	PER	Sandy
2	LOC	Maryland	1	ORG	AP
1	LOC	NY City	1	ORG	CNN
1	LOC	New Hampshire	1	ORG	Coriolis Effect
5	LOC	New Jersey	1	ORG	Little Ferry
7	LOC	New York	1	ORG	MTA
2	LOC	New York City	1	ORG	Metropolitan Transit Authority
1	LOC	New York Stock Exchange	1	ORG	Moonachie
1	LOC	New York University	1	ORG	National Weather Service
1	LOC	Ohio	1	ORG	New York Stock Exchange
1	LOC	Queens	1	ORG	Newark Liberty
1	LOC	Teterboro	1	ORG	Tisch Hospital
4	LOC	US	1	ORG	Trams
1	LOC	Washington DC	1	ORG	US Department of Energy

Tables 6 and 7 explore the fifth strongest topic of this news article, "physics", showing a collection of the strongest individual Wikipedia articles on this topic, and strongest features of this topic.

The Finnish language Wikipedia turned out to be far less extensive than the English one. Instead, we used a collection of 73 000 news articles from the Finnish News Agency (STT). Generally, the text in this material is of good quality, but there are some limitations: sports news are dominating and there are very few information technology related news (no Apple, Google, Facebook, Twitter, etc.). The STT news used here date from the years 2002-2005 including also 5000 news from February 2013. For POS tagging the STT news we used a commercial morphological parser, FINTWOL by Ling-Soft Ltd., and for NER tagging we created lists of NER tagged words to which we compared single and groups of POS tagged and lemmatized words. As an example for Finnish, Tables 8 and 9 present results based on an article about the re-election of Giorgio Napolitano as the president of Italy (*Talouselämä*, 22 April 2013).

Table 6: The strongest individual Wikipedia articles for the topic "Physics"

Terahertz time-domain spectroscopy
List of materials analysis methods
Fiber laser
Cryogenic particle detectors
Varistor
Neutron generator
Laser ultrasonics
Optical amplifier
Thyristor
Electric current
Neutron source
Voltage-regulator tube
Switched-mode power supply
Gas-filled tube
Isotopes of plutonium
Superconducting magnet

Table 7: The strongest features for the topic "Physics"

NE-LOC	US, Europe, Chernobyl, Hiroshima, Earth.
NE-MISC	X-ray, Doppler, CO2, °C, CMOS, Fresnel.
NE-ORG	CERN, IPCC, IAEA}.
NE-PER	Maxwell, Edison, Gibbs, Watt, Richter, Einstein, Rutherford, Faraday, Bohr}.
POS-JJ	nuclear, electrical, magnetic, liquid, thermal, atomic, mechanical, solid}.
POS-NN:	energy, power, system, gas, material, temperature, pressure, air, effect, frequency, wave, field, heat, particle, unit, process, signal, mass, device, surface, circuit, light.
POS-RB:	relatively, extremely, slowly, fast.
POS-VB:	produce, require, cause, measure, reduce, increase, generate, allow, apply, create.

Table 8: The five strongest topics for a Talouselämä, 22 April 2013 article (English translation in parenthesis)

Number	Weight	Topic Name
1	0.1014	Vaalit (elections)
2	0.0567	Kansainvälinen konflikti (international conflict)
3	0.0497	Sää (weather)
4	0.0438	Aseellinen selkkaus (armed conflict)
5	0.0434	Tuloneuvottelut (income negotiations)

Table 9: The Named Entities for the Talouselämä, 22 April 2013 article

Freq.	Type	Value
2	MISC	presidentti (president)
1	MISC	radikaali (radical)
2	LOC	Italia (Italy)
1	LOC	maa (country)
2	PER	Napolitano
1	ORG	hallitus (government)
2	ORG	parlamentti (parliament)

Figure 2 shows all the fifty topics obtained for the *Talouselämä*, 22 April 2013 article. The highest peak is the strongest topic "president". The number of Named Entities for the example in Finnish is much smaller than that in English. The state of the art NER taggers for Finnish are not as evolved as the taggers for English.

The overall results are, in fact, better for the BBC article; there are more NER tagged words and the strongest topics correspond better to the semantic contents of the article.

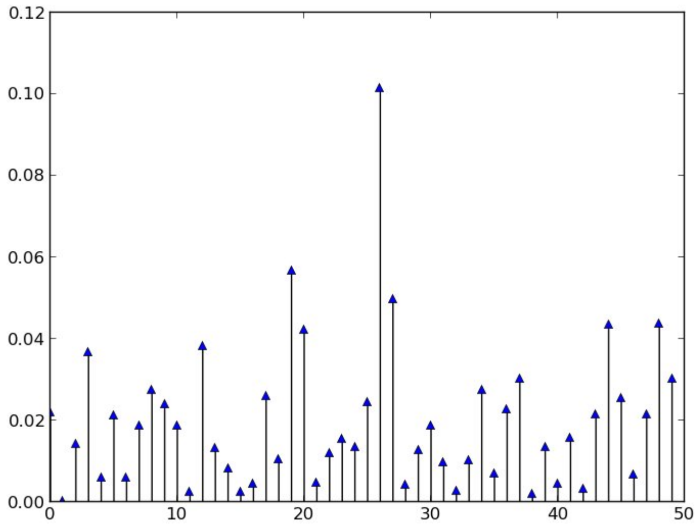

Figure 2: The fifty strongest topics for one news article projected against model created with STT news data. The horizontal axis shows the number of the topic and the vertical axis shows normalized weight of the topic

However, the results using the model created with STT data are far better than those created with the Finnish Wikipedia. This is demonstrated in Figure 3 where the strongest topics do not as strongly rise above the rest and, furthermore, the five strongest topics are mostly not significant: Finnish politics, philosophy and religion, natural sciences, computer games, and banks and monetary policies. This shows that the corpus and named entity data used to create the model is sufficiently extensive and of good quality.

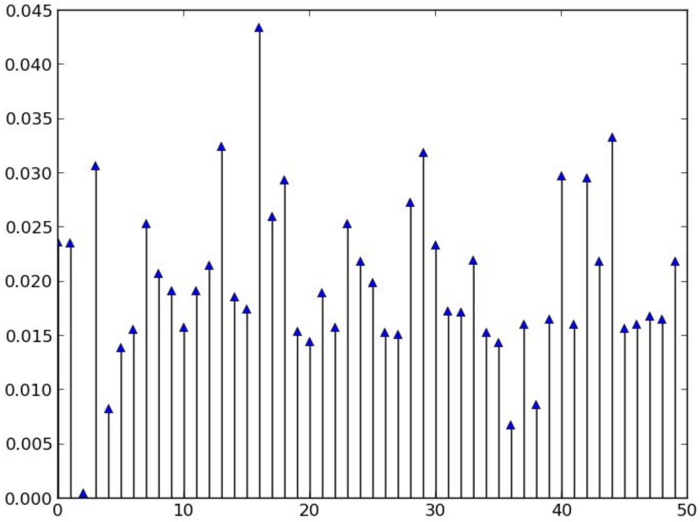

Figure 3: The fifty strongest topics for one news article projected against a model created using the Finnish Wikipedia

As another source for the corpus in Finnish we used the free newspaper sheet *Metro*. For POS tagging of *Metro* news we used the Open Source Morphology for Finnish, OmorFi (Lindén et al., 2011), and for NER tagging we used a combination of OMorFi and our own POS tagging version created for STT news.

3.2 Word association analysis

3.2.1 Diverging word clouds

In this section we present some results of the Software Newsroom applications that use word association analysis as their basis. As before, for single document experiments we use the English Wikipedia as the background corpus and a story from BBC, "Google tests balloons to beam internet from near space" (Kelion, 2013), as the foreground document that we are interested in.

Given a document *d* and a word *w*, a specification for the corresponding diverging association cloud is directly obtained from the foreground associations of the document: take the top *n* words associated with *w* in the foreground and position them in the word cloud according to their weights. Figure 4 illustrates the idea using the document on Google balloons.

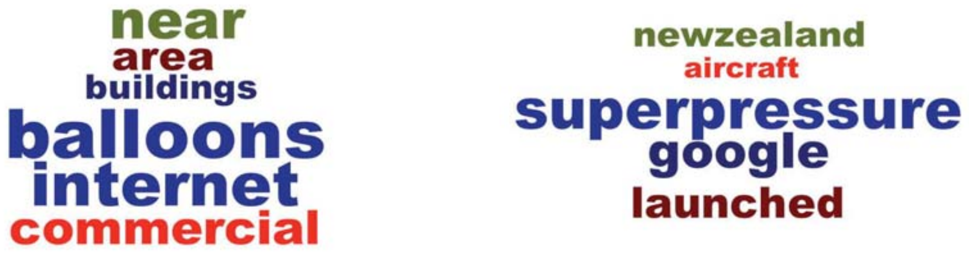

Figure 4: The diverging word cloud created from the foreground associations of the news story "Google tests balloons to beam internet from near space". The search term for the left diverging cloud is 'google' and for the right diverging word cloud the search term is 'balloons'. We used an internet tool (Word Clouds for Kids, 2012) for generating the word clouds

These association clouds give a good idea of what the document could be about. Such word clouds could also be implemented in an interactive manner: as the user clicks on a word in the cloud, the selected word becomes the next search term and, correspondingly, all divergent clouds are re-rendered for all documents using the new focus. A drawback of this method is that it takes time to get used to the fact that the word cloud is conditioned on the search term and thus interpreting the results could be non-intuitive for novice users. In order to alleviate the problem, it would be possible to also present the search term together with the words but currently we do not have a clear idea of how to present this in an intuitive manner.

3.2.2 Summary generation

For our experiments with the summary generation algorithm presented earlier, we collected four news stories on the same topic from different news sources - BBC: "Up, up and away: Google to launch Wi-Fi balloon experiment" (Kelion, 2013), National Geographic: "Google's Loon Project Puts Balloon Technology in Spotlight" (Handwerk, 2013), ARS Technica: "Google's balloon-based wireless networks may not be a crazy idea" (Brodkin, 2013), CNN: "Google tests balloons to beam internet from near space" (Smith-Spark, 2013).

For the background associations we used Wikipedia news stories and the foreground associations were calculated using the respective news stories. We then applied the algorithm which we described on this set of data. The total number of words in all the four documents was 3696.

The first four sentences, containing a total of 120 words, returned by the algorithm were the following:

- Google is reportedly developing wireless networks for sub-Saharan Africa and Southeast Asia that would combine a technology well established for such purposes (TV White Spaces) with one that's a bit more exotic - balloons that transmit wireless signals. - *ARS Technica*
- Project Loon balloons are made of plastic just 3 mm (0.1in) thick, another Orlando-based firm, World Surveillance Group, sells similar equipment to the US Army and other government agencies. - *BBC*
- It has been working on improving connectivity in the US with Google Fiber and bringing the internet to underserved populations overseas through White Spaces networks. - *ARS Technica*
- A company called Space Data makes balloon-based repeater platforms for the US Air Force that "extend the range of standard-issue military two-way radios from 10 miles to over 400 miles." - *ARS Technica*

The application of the algorithm yields promising results. Our next goals are improving and evaluating the current method. It is important to note here that the way the extracted sentences are presented to the user is also a very important aspect. For instance, consider the third sentence which has a co-reference resolution problem (i.e., the sentence starts with "it" and we do not know what "it" is). In such cases it makes sense to present consecutive sentences together in the summary regardless how they are ordered by the algorithm. In some cases this could help to overcome the co-reference resolution problem. It is also possible to provide some context to the user, for instance, when the user's cursor hovers over an extracted sentence, the sentences which are before and after it in the news story can be shown.

4. Discussion

Application of MPCA seems to work well for news search by topic analysis. It is likely that also other variants of probabilistic modeling perform well for news identification. Our second approach, association analysis, also clearly enhances the effectiveness of the "news nose". A question then arises, whether other methods could be effective as well, or even better than the adopted approaches.

In contrast to statistical methods such as PCA, *cluster analysis* can best be seen as a heuristic method for exploring the diversity in a data set by means of pattern generation (van Ooyen, 2001). Cluster analysis may be applied for finding similarities and trends in data (described using the common term *pattern recognition*). An example of cluster analysis is the *expectation maximization* (EM) algorithm, which has recently been applied to astronomical data for identifying stellar clusters from large collections of infrared survey data (Solin, Ukkonen and Haikala, 2012). Cluster analysis has also been used in, e.g., market research within a more general family of methodologies called *segmentation methods*. These can be used to identify groups with common attitudes, media habits, lifestyle, etc. Cluster analysis is probably less well suited for news search than probabilistic models like MPCA, since the semantic contents of articles that contain more than one topic are not resolved by cluster analysis (e.g., Newman at al., 2006), while probabilistic modeling clearly performs well in such cases provided that the corpus and named entity data used for the model creation are sufficiently extensive. This will result in only a small number of unrecognized words that cannot be tagged, and thus a high resolving power of topics and named entities.

Supervised learning methods divide objects such as text documents into predefined classes (Yang, 1999). Cluster analysis and PCA are data driven methods which can extract information from documents without *a priori* knowledge of what the documents may contain (Newman et al., 2006), and topic categorization (i.e., a topic model) is created by the algorithm without rules or restrictions on the contents of a topic, which is why such methods are called *unsupervised learning*. Obviously supervised learning is poorly suited for news search from arbitrary textual data, since the topics of potential news in the material cannot be predicted, and it is thus impossible to recognize new emerging topics.

A further, more advanced analysis of complex data may incorporate the use of *semantic networks*. Methods of this category are *Traditional* and *Improved Three-Phase Dependency Analysis* (TTPDA, ITPDA). These algorithms have been applied to recognition of semantic information in visual content and they use Bayesian networks to automatically discover the relationship networks among the concepts. These methods can be applied, for example, to automatic video annotation. (Wang, Xu and Liu, 2009).

In this paper, we have mainly interpreted the associations on a single association level rather than as a network. But these associations, both background and foreground, can also be seen as a kind of semantic network where words are nodes and the edges represent the associations. Analyzing the background associations as a network might give interesting results in automatic word domain discovery or for finding interesting subnetworks that connect two words. The same applies for the foreground associations, which might provide interesting inference and application possibilities when interpreted as word networks and used as such. Thus, in the future, our models and methods could be improved in their accuracy. More efficient, scalable algorithms could be designed and, perhaps more interestingly, additional novel applications could be invented with help of the background and foreground models, especially in the broad areas of information browsing and retrieval.

Considering the topic model and data used for the training, our experiments indicate that the comprehensiveness, quality, and also the semantic similarity of the text corpus and named entity data with the data under analysis are critical to the effectiveness of the search algorithm. This is of course obvious, but poses a challenge for automated news search since language evolves and the language used in, e.g., social media that obeys no standard rules diffuses with an increasing speed to various media channels. Should we accept this and modify the models and additionally also adopt slang in the presentation of news, or try to force the users to educate themselves in order to write in decent standard language also in social media?

An aspect of crucial importance in (automated) news search is the quality of the data. The internet is full of hoaxes and distorted information, and finding assurance for the reliability of potential news may sometimes be challenging, and will require too much time.

This may lead to that the potential news becomes yesterday's news or that it is published by a competitor before sufficient background information is found. The Software Newsroom should therefore trace all possible metadata on the sources, time, places, people, and organizations associated with the creation of the information found by automated means. While this cannot rely merely on software, automation can be used to significantly improve the effectiveness and speed of the process.

5. Conclusions

We have developed and applied methods for automated identification of potential news from textual data for use in an automated news search system called Software Newsroom. The purpose of the tools is to analyze data collected from the internet and to identify information that has a high probability of containing news. The identified potential news information is summarized in order to help understanding the semantic contents of the data and also to help in the news editing process.

Two different methodological approaches have been applied to the news search. One method is based on topic analysis which uses MPCA for topic model creation and data profiling. The second method is based on association analysis that applies LLR. The two methods are used in parallel to enhance the news recognition capability of Software Newsroom.

For the topic mining we have created English and Finnish language corpora from Wikipedia and several Finnish language corpora from Finnish news archives, and we have used bag-of-words presentations of these corpora as training data for the topic model. We have made experiments of topic analysis using both the training data itself and arbitrary text parsed from internet sources. The selected algorithmic approach is found to be well suited for the task, but the effectiveness and success of news search depends strongly on the extensiveness and quality of the training data used for the creation of the topic model. Also, semantic similarity of the target text with the corpus used for the model creation generally improves the search effectiveness. The large difference between the language commonly used in user-created internet content and standard language poses a challenge for news search from social networks, since a significant part of the language is not recognized by the part-of-speech and name entity taggers. A simple solution for this would be a translator that would preprocess the unknown slang words, turning them into standard language. Another would be a slang-based corpus. The latter has the disadvantage that the resulting raw news material would be composed of slang and it would have to be translated into standard language before publishing. Thus, our plan is to collect a small dictionary of the most common words used in social media and use them for further experiments on social media.

In the association analysis we have used a methodology for detecting novel word associations from a set of documents. For detecting novel associations we first used the background corpus from which we extracted such word associations that are common. We then collected the statistics of word co-occurrences from the set of documents that we are interested in, looking for such associations which are more likely to appear in these documents than in the background.

We also demonstrated applications of Software Newsroom based on association analysis - association visualization as a graph, diverging (association) clouds which are word clouds conditioned on a search term, and a simple algorithm for text summarization by sentence extraction. We believe that the background-foreground model has significant potential in news search. The simplicity of the model makes it easy to implement and use. At the same time, our experiments indicate great promise in employing the background-foreground word associations for different applications.

The combination of the two methods has not yet been implemented. This is in our plans for the near future. and the application of both methods on social media using a pre-translator of social media language is underway. Potential future work also includes experiments on automated news generation and application of our methods for other purposes, e.g., improvement of recommendations.

Acknowledgements

We are grateful to Jukka Mauno from MTV3 for raising the idea of Software Newsroom. We acknowledge Tekes, the Finnish Funding Agency for Technology and Innovation, and Tivit Oy for financing this research. Especially we wish to thank the Focus Area Director of the Next Media research programme, Eskoensio Pipatti, for his support. This work has also been supported by the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland and Hecse, Helsinki Doctoral Programme in Computer Science - Advanced Computing and Intelligent Systems. Finally, we thank the referees for valuable comments.

References

Agrawal, R., Imieliński, T. and Swami, A., 1993. Mining association rules between sets of items in large databases. *ACM SIGMOD Record.* 22(2), pp. 207-216

Allan, J., 2002. *Topic Detection and Tracking: Event-based Information Organization.* Dordrecht: Kluwer Academic Publishers.

Berry, M.W., Dumais, S.T. and O'Brien, G.W., 1994. Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review,* 37(4), pp. 573-595

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, pp. 993-1022

Brodkin J. (2013). Google's balloon-based wireless networks may not be a crazy idea. *ARS Technica*, June 2, 2013. [Online] Available at: <http://arstechnica.com/information-technology/2013/06/googles-balloon-based-wireless-networks-may-not-be-a-crazy-idea/>. [Accessed 26 June 2013]

Buntine, W. and Jakulin, A., 2004. Applying discrete PCA in data analysis. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (UAI2004), pp. 59-66

Buntine, W. and Jakulin, A., 2006. Discrete component analysis. *Subspace, Latent Structure and Feature Selection, Lecture Notes in Computer Science.* Vol. 3940, pp. 1-33

Church, K.W. and Hanks, P., 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), pp. 22-29

Comintelli, 2013. [Online] Available at: <http://www.comintelli.com/Company/Press-Releases/Esmerk-launches-new-current-awareness-platform-Esm> [Accessed 31 October 2013]

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

Dumais, S. T., Furnas, G.W., Landauer, T.K., Deerwester, S. and Harshman, R., 1988. Using latent semantic analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in computing systems,* pp. 281-285

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1), pp. 61-74

Hakulinen, A., Vilkuna, M., Korhonen, R., Koivisto, V., Heinonen, T. R., and Alho, I., 2004. *Iso suomen kielioppi.* Helsinki: Suomalaisen Kirjallisuuden Seura (Available online at http://scripta.kotus.fi/visk/etusivu.php)

Handwerk, B. (2013). Google's Loon Project Puts Balloon Technology in Spotlight. *National Geographic*, June 18, 2013. [Online] Available at: <http://news.nationalgeographic.com/news/2013/06/130618-google-balloon-wireless-communication-internet-hap-satellite-stratosphere-loon-project/>. [Accessed 26 June 2013]

Harris, Z., 1954. Distributional Structure. *Word*, 10(23), pp 146-162

Hofmann, T., 1999. Probabilistic Latent Semantic Indexing, *Proc. 22nd Annual International SGIR Conference on Research and Development in Information Retrieval*, pp. 50-57

Hori, C. and Furui, S., 2003. A new approach to automatic speech summarization. *IEEE Transactions on Multimedia*, 5(3), pp. 368-378

Huang, S., Peng, X., Niu, Z. and Wang, K., 2011. News topic detection based on hierarchical clustering and named entity. *7th International Conference on Natural Language Processing and Knowledge Engineering.* pp. 280-284

Hynönen, T., Mahler, S. and Toivonen, H., 2012. Discovery of novel term associations in a document collection. *Bisociative Knowledge Discovery, Lecture Notes in Computer Science,* Vol. 7250, pp. 91-103

Karlsson, F., 1990. Constraint grammar as a framework for parsing running text. *Proceedings of the 13th Conference on Computational Linguistics*, Vol. 3., pp. 168-173

Kelion L. (2013). Google tests balloons to beam internet from near space. *BBC*, June 15, 2013. [Online] Available: <http://www.bbc.co.uk/news/technology-22905199>. [Accessed 26 June 2013]

Kimura, M., Saito, K. and Uera, N., 2005. Multinomial PCA for extracting major latent topics from document streams, *Proc. 2005 IEEE International Joint Conference on Neural Networks*, Vol. 1, pp. 238-243

Knight, K. and Marcu, D., 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1), pp. 91-107

Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A. and Silfverberg, M., 2011. HFST-Framework for Compiling and Applying Morphologies. In: Mahlow, C. and Piotrowski, M., eds.(2011). *Systems and Frameworks for Computational Morphology. Communications in Computer and Information Science*, Vol. 100. Berlin-Heidelberg: Springer. pp. 67-85.

Lindén, K. and Pirinen, T., 2009. Weighted finite-state morphological analysis of Finnish compounds. In: Jokinen, K. and Bick, E., eds.(2009). *Proc. Nordic Conference of Computational Lingustics*. Odense: NEALT

McDonald, D.D., 1996. Internal and external evidence in the identification and semantic categorization of proper names. In: Boguraev, B. and Pustejovsky, J., eds. (1996). *Corpus Processing for Lexical Acquisition*. Cambridge, MA: MIT Press. pp. 21-39

Meltwater, 2013. [Online] Available at: <http://www.meltwater.com/products/meltwater-buzz-social-media-marketing-software/> [Accessed 31 October 2013]

Miller G.A., 1995. WordNet: A Lexical Database for English, *Communications of the ACM,* 38(11), pp. 39-41

Nadeau, D. and Sekine, S., 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1), pp. 3-26

Newman, D., Chemudugunta, C., Smyth, P. and Steyvers, M., 2006. Analyzing Entities and Topics in News Articles using Statistical Topic Models. *Lecture Notes in Computer Science*, Volume 3975, pp. 93-104

van Ooyen, A., 2001. Theoretical aspects of pattern analysis. In: L. Dijkshoorn, K. J. Tower, and M. Struelens, eds. *New Approaches for the Generation and Analysis of Microbial Fingerprints*. Amsterdam: Elsevier, pp. 31-45

Padró, L., Reese, S., Agirre, E. and Soroa, A., 2010. Semantic Services in FreeLing 2.1: WordNet and UKB. *Proceedings of the Global Wordnet Conference 2010*.

Pearson, K., 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2(11), pp. 559-572

Ratinov, L. and Roth, D., 2009. Design Challenges and Misconceptions in Named Entity Recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147-155

Salton, G., 1991. Developments in Automatic Text Retrieval. *Science*, Vol 253, pp. 974-979

Shen, D., Sun, J.T., Li, H., Yang, Q. and Chen, Z., 2007. Document summarization using conditional random fields. *Proceedings of the 20th international joint conference on Artifical intelligence,* Vol. 7, pp. 2862-2867

Smith-Spark L. (2013). Up, up and away: Google to launch Wi-Fi balloon experiment. *CNN*, June 15, 2013. [Online] Available at: <http://www.bbc.co.uk/news/technology-22905199>. [Accessed 26 June 2013]

Solin, O., Ukkonen, E., and Haikala, L., 2012. Mining the UKIDSS Galactic Plane Survey: star formation and embedded clusters, *Astronomy & Astrophysics*, Volume 542, A3, 23 p

Toivonen H., Gross, O., Toivanen J.M. and Valitutti A., 2012. Lexical Creativity from Word Associations. *Synergies of Soft Computing and Statistics for Intelligent Data Analysis. Advances in Intelligent Systems and Computing.* Vol. 190*,* pp. 17-24

Wang, F., Xu, D. and Liu, J., 2009. Constructing semantic network based on Bayesian Network. *1st IEEE Symposium on Web Society*, pp. 51-54

Word Clouds for Kids, 2013. [Online] Available at: <http://www.abcya.com/word_clouds.htm>. [Accessed 26 June 2013]

Yang, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, Vol 1, pp. 67-88

Yeh, J.Y., Ke, H.R., Yang, W.P. and Meng, I., 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management*, 41(1), pp. 75-95