Towards Topics-based, Semantics-assisted News Search

Martin Voigt
TU Dresden
01062, Dresden, Germany
martin.voigt@tu-dresden.de

Michael Aleythe, Peter Wehner
Fink & Partner Media Service GmbH
01309, Dresden, Germany
{michael.aleythe,peter.wehner}@finkundpartner.de

ABSTRACT

Identifying upcoming topics from a news stream is a challenging and time consuming task for editors since they have to recognize proper keywords, actively search with them, and need to browse the located media assets. To this end, our goal is to enhance an existing newsroom environment to automatically detect upcoming global and regional topics which are suggested for editors further work. To understand the impact of a topic, we provide its evolution over the time and the relations to other subjects as helpful indicators. To achieve our goals, we designed and prototypically implemented an automatic, semantics-based workflow which heavily relies on non-ambiguous named entities extracted from the media assets. Further, we discuss the challenges encountered and point to proper solutions for building your own enterprise-scaled semantics-based application.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks

General Terms

Algorithms

Keywords

topic recognition, trend analysis, news search, semantics

1. INTRODUCTION

Within the daily production process of print and online media the journalist is hindered by a massive amount of media assets like articles, photos, and videos. In a mid-sized publishing house solely 2000 textual assets reach the newsroom from different agencies. Thus, it becomes pretty hard to investigate upcoming global and local topics and their development over the time. Moreover, it is also challenging to identify relevant assets for a subject in the news stream and maybe in their media archive. Existing tools became more and more obsolete as the user has to know explicitly what to search for. Furthermore, they are mostly keyword-based what makes it hard to search about complex relation of entities and their occurrence in articles.

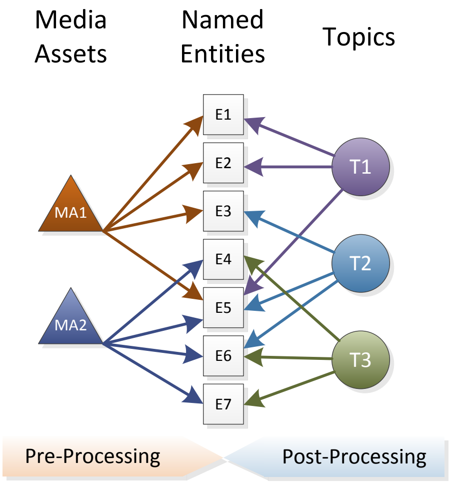


Figure 1: Schema of using named entities from media assets to recognize topics.

Therefore, our vision is to investigate a tool which automatically identifies and represents upcoming or current *topics* within the news stream. The issues – regardless if they have a global or regional scope – are presented in a ranked list to the journalist according their currentness and weighting against other topics. Further, he could simply browse all related media assets, pick them up, evaluate their usefulness and finally write his own article. The benefits for the publishing houses are obvious: the journalists save time during their investigation by identifying only the relevant assets. This yields to media with an higher quality, maybe in less time.

Fig. 1 gives an overview of our approach. We try to identify named entities within media assets in a so-called preprocessing step, which is elaborated in Sect. 4.1. In a postprecessing, we try to identify frequent combinations of two or more named entities. The example in Fig. 1 comprises three topics which are composed of three named entities in each case. An entity could be part of multiple subjects, e. g., E5 and E6 in our example.

WIMS '13, June 12-14, 2013 Madrid, Spain
Copyright 2013 ACM 978-1-4503-1850-1/13/06 ...$15.00.

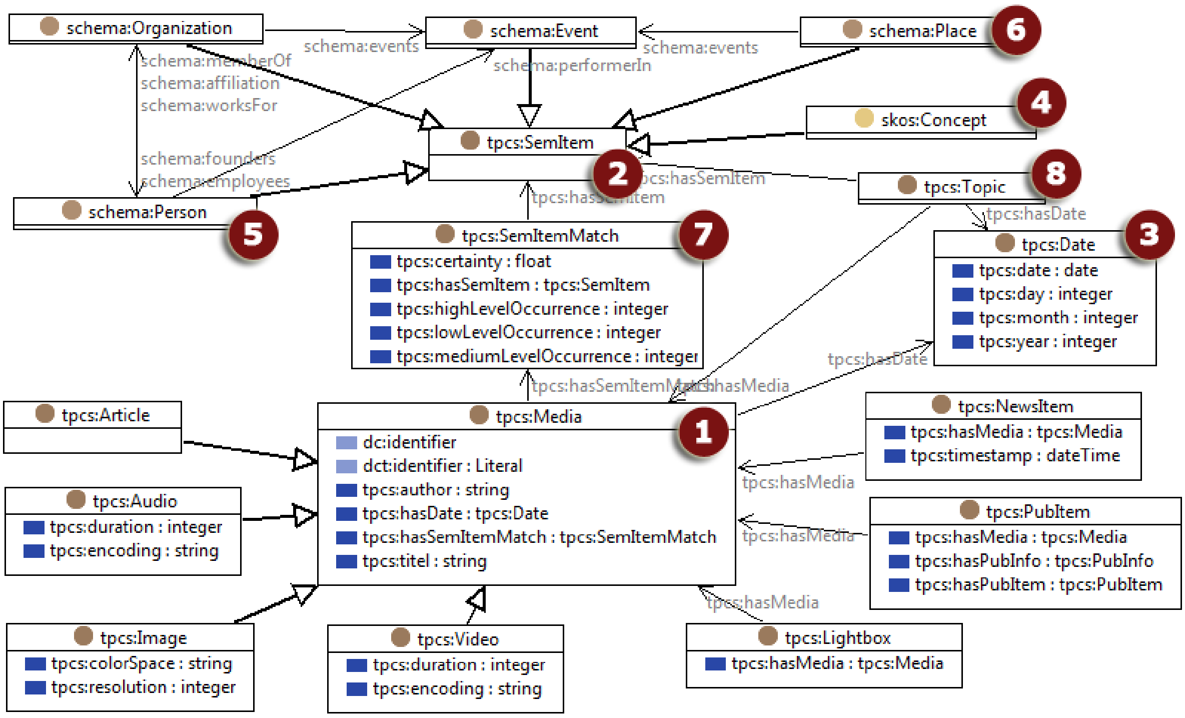


Figure 2: Excerpt of the topics schema.

To reach our goals a lot of research and practical problems need to be solved. The most high-level ones are: 1) the semantic modeling of topics including their relation amongst each other, 2) the extraction and disambiguation of information which are the foundation of the topics, 3) the assessment and ranking of a topic using quantitative as well as time-related measures, and 4) a scalable architecture which seamlessly integrates in an existing newsroom environment and implements our vision of a topic-based news search.

After sketching the related work in Sect. 2, we give an overview of our current semantic model in Sect. 3. Afterwards, we present the workflow of our system (Sect. 4) which allows for analyzing incoming media assets through presenting the topics from the news stream. Furthermore, we point to and discuss the key challenges identified (Sect. 5) since we implemented the foundational parts of the process and evaluated them using a big real-world dataset. Some of them may also underline the existing research agenda of the semantic web community.

2. RELATED WORK

In the following, we give a brief overview of other research and commercial approaches to assist users to identify topics and trends within news.

The *Europe Media Monitor* [2] is a research project founded and developed by the European Commission. It collects, aggregates, and classifies news from more than 60 news portals in different languages. Based on extracted named entities it derives clusters and trends of news articles. As we follow the same approach, we are not focusing on the end user reading news but the journalists. They need to create articles using the more “raw” newswire texts and other media assets like photos and videos. Besides global news we also address the regional content, thus, persons like the mayors of small cities or local football players are maybe more interesting than Angela Merkel. Finally, we are building on semantic technologies and making use of the knowledge provided through the Linked Data Cloud [5] what allows for data enrichment, reasoning, and disambiguation.

We also identified some commercial tools like *APA DeFacto*[1] and *dpa-Agenda*[2] which address the expert users who need to search for and work with media assets delivered first and foremost by news agencies. Their search results are grouped by subjects but the recognition of upcoming and current topics is missing. Especially dpa Agenda is working with manual created topics which are following their own news coverage. The development of a subject over the time is not available as well. Further, the tools miss to leverage the information from the Linked Data Cloud.

We should also mention other projects like NEPOMUK [4] or Twitcident [1]. Although they address parts of our workflow proposed in Sect. 4, e. g., text analysis, semantic information modeling and search, our approach is different in detail due to the varying scopes. We are focusing on the news media domain with a broad spectrum of categories, e. g., politics, sports, or culture, and a high throughput of media assets per day. In contrast to some research projects, we had a real business use case which requires a high precision and a scalable architecture. Furthermore, in NEPOMUK, Primal[3] and others a topic is not know a priori or not a combination of semantical disjunct entities. In the end, we strive for a system which reports new topics to its user, hence, he has not to search for it.

[1] http://www.apa-defacto.at
[2] http://goo.gl/ik5Ua
[3] https://primal.com/

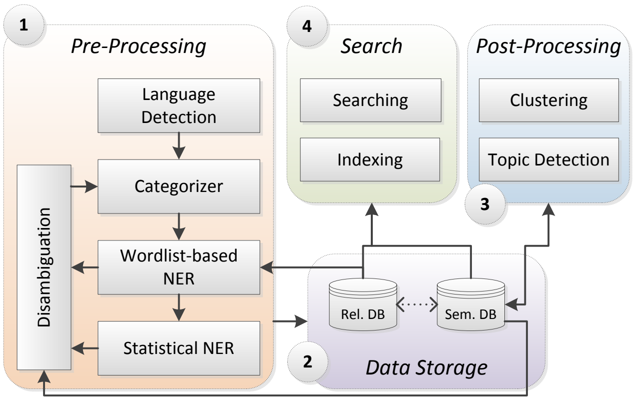


Figure 3: Overview of the Topic/S workflow

3. SEMANTIC MODEL

The foundation of our concept is a semantic model to describe the media assets and their life-cycle within our system. Since it is our first project employing semantic technologies for information management, we followed Hendlers idea of “A Little Semantics Goes a Long Way”[4] and modeled it iteratively by using RDFS. Further, we do not aim to reinvent the wheel, thus, we are re-using existing vocabulary or schemata like SKOS, IPTC NewsCodes[5], or schema.org.

Fig. 2 gives an overview of the most important parts of the semantic model. Its core is the interlinking of *Media* assets ① with “semantic items”, called *SemItems* ②. A media asset, which arrives within the system at a specific *Date* ③, could be a news *Article*, an *Image*, an *Audio* or *Video* file. Using our pre-processor (cf. Sect.4), we identify the category of the item, which is modeled as a SKOS *Concept* ④ by using IPTC NewsCodes. Furthermore, we extract and annotate named entities, e. g., *Persons* ⑤ or *Places* ⑥, from the texts or metadata. Since it is not possible to define N-ary relations using the current semantic web standards from the W3C and we need to specialize the relations between *Media* and *SemItems*, we defined the so-called *SemItemMatch* concept ⑦. It allows us to specify the certainty of a category but also to define how often a named entity is enclosed, e. g., within the text of an *Article* (*tpcs:lowLevelOccurrence*). Finally, we are modeling *Topic* clusters ⑧ which are linking *SemItems* with *Media*.

4. TOPIC/S WORKFLOW

Fig. 3 provides an overview of the Topic/S workflow which is developed as a module of a newsroom environment. It comprises four main steps: the extraction of semantic information contained in the assets ①, their storage ②, the topic identification ③, and finally a module to allow for their search ④. In the following, we describe the functionality of each step more detailed.

4.1 Pre-Processing

The media assets, currently textual articles from news agencies, are passed to the pre-processor ①. The first step is to detect the language of the text. Currently, we are importing German and English texts thus a self-defined set of rules using common statistics like the occurrence of umlauts or the number of language-typical words, e. g., ‘of’ and ‘for’ in English texts, helped us to reach a precision of 99%.

Since every news agency employs its own classification of the media assets, we are categorizing them on our own to have an homogeneous systematization. As this step requires specifically trained models the correct categorizer is chosen according to their language and agency. After some studies, we decided on LingPipe[6], its NaiveBayes classifier, and the IndoEuropeanTokenizer. Table 1 gives an impression about the precision for different agencies trained with at least 1.000 or 2.000 articles. It shows that smaller, domain-specific news agencies like KNA and EPD perform worse in contrast to the big players like DPA or Reuters. The reason is that their texts are biased according their domain, e. g., for KNA and EPD the texts are tailored to an religious audience, what makes it harder to identify a category. Furthermore, through the statistics in Table 1 we understand that more training would not enhance the precession significantly.

Agency	Precision 1.000 articles	Precision 2.000 articles
KNA	80,3%	81,9%
DPA	94,4%	94,2%
EPD	80,3%	83,5%
Reuters	90,8%	90,4%
OTS	93,5%	93,8%

Table 1: Evaluation of our categorization approach for different news agencies using LingPipe trained with at least 1000 or 2000 articles.

We integrated two components to detect named entities contained in the text. The reasons why we are using two approaches concurrently are discussed in-depth in Sect. 5. The first is a wordlist-based approach implemented by using LingPipe as well which reaches a much higher precision than statistical one. The lists are semi-automatically retrieved from the YAGO2 [6], the German DBpedia[7], Freebase[8], and GeoNames[9]. They comprise the names – if possible in different spellings – and an unique URI. Table 2 gives an overview of the number of entries of our four distinct lists. At present, we face the task to identify important keywords besides the well-known named entities, which help us to create more meaningful topics, e. g., war or peace. Therefore, extract the topmost used nouns - less our named entities – from a set of 1.000 articles in their canonical form and rated them manually if they are applicable or not.

The second component is statistics-based and allows for recognizing concepts which are not include in the word lists. We use the Stanford Natural Language Processing Tools[10]. Again, the trained model is selected in compliance with the discovered language. The results are stored separately, thus, a distinguished user has to check them and may add new named entities to the word lists if applicable.

The determined entities are often ambiguous. A good

[4] http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html

[5] http://www.iptc.org/site/NewsCodes/

[6] http://alias-i.com/lingpipe/

[7] http://wiki.dbpedia.org/

[8] http://www.freebase.com/

[9] http://www.geonames.org/

[10] http://nlp.stanford.edu/software/

Semantic Item	Number of Entries
Persons	590.828
Organizations	63.262
Places (Germany)	89.672
Keywords	1.329

Table 2: Overview of the number of entries in our keyword lists for the NER.

example is the word “Golf”[11]. Therefore, we are currently developing a dedicated component to allow for their disambiguation. At first place, it uses our internal knowledge base ② with the already extracted semantic items which occur often in specific relations, e. g., golf and sport or golf an car. In a second step, we are developing an algorithm to extract and use entity relations from DBpedia if our model is too imprecise.

4.2 Data Storage

After the information has been extracted, the results are stored within a storage layer ②. Therefore, we conducted a triple store benchmark on multiple criteria [12] and observed that comprehensive tests with real-world data are necessary to detect anomalies of the stores, e. g., the different results of the RDFS inference or the varying performance of SPARQL 1.1 UPDATE queries. Unfortunately, we could not recommend any triple store in general as no store could win in all fields. Thus, the selection strongly depends on your specific project requirements.

We rely on Oracle 11gR2[12] which allows us to easily combine a relational database with a semantical one. Since the first holds general information about the media assets, e. g., the text of an article, the latter manages the information extracted during the pre-processing by using the semantic model presented in Sect. 3. Such an integrated concept, which is also provided by Virtuoso[13], has several advantages: First, the huge number of articles can be handled by the faster relational database and will not blow up the triple store. Second, the combination of SQL and SPARQL statements allows for the integration of the relational and semantical world within one query and further, it lowers the barrier of using SPARQL for expert SQL users. Third, the usage of existing mining features like spatial data mining is available for RDF data, too.

For our prototypical tests we are using newswire articles as well as archived material from the Main-Post[14]. Both are continuously imported into the database to simulate the growing amount of data a productive system has to handle. Currently, we are managing about 200.000 articles, thus, ca. 10.3 million triples are generated through pre-processing ① and stored in the semantical database ②.

4.3 Post-Processing

The post-processing ③ allows for reducing the dataset and to investigate what are upcoming news topics. Currently, it is our most important task to develop an algorithm to identify appropriate topic clusters. Like introduced in Sect. 1 and in Fig. 1, our conceptual idea is to identify frequent combinations of two or more semantic items which build a topic. Their importance is reflected by the number of assigned media assets per day. On considering a time period, we are able to see the trend of a topic.

The two main prerequisites for our scenario are that the number of topic clusters could not be predefined and that clusters could overlap. After investigating the related work, we decided on hierarchical clustering, especially the agglomerative approach which is faster and easier to implement as the divisive one [7]. Our current algorithm works like following:

1. Create a triangular matrix with every article as column and row.

2. Calculate the similarity between every article using the *Dice coefficient* where we use the linked *SemItems*, thus, persons, organizations, places, and keywords.

3. To reduce the number of calculations, we implemented a “pre-clustering” method to merge all articles with a similarity of 1.0. This eliminates for example duplicate articles which were used in different regional issues of a news paper.

4. Identify and merge the articles with the highest similarity and store their child nodes.

5. Now, the similarity of all other articles to the new cluster is computed using the *Complete Linkage* method. After some evaluations we found that it presents the topics best. Alternatives were *Single* or *Average Linkage*. Write the new value into the matrix.

6. The last two steps are repeated until all similarity values are zero what is different from the traditional approach where the algorithms runs until a complete cluster is created. This is not required in our case.

7. We filter the result list with regard to two characteristics. First, we remove all topics with only a few articles and many semantic items since they are very special. Second, we discard all topics with many articles and less SemItems as they are very general.

8. Finally, the results are stored within our triple store where every topic is linked to a specific date (Sect. 3). We store only one “topic model” per day, hence, the updates within a day may overwrite existing calculations.

At present, we try to configure our algorithm to enhance the topic quality and performance. But in general, we are a kind of satisfied since one run lasts only about 50 s for ca. 2.000 articles of a day using a simple desktop computer. Table 3 gives an impression of the topic clusters generated for one day using our test dataset.

Furthermore, we target to reduce the information overload of the editors by using spatial clustering, thus to assemble news of a specific region. Therefore, we already augmented

[11] http://en.wikipedia.org/wiki/Golf_%28disambiguation%29

[12] http://www.oracle.com/technetwork/database/

[13] http://virtuoso.openlinksw.com/

[14] http://www.mainpost.de/

Rank	Articles	SemItems	Similarity	"Name"	Is Topic?
1	27	3	0.235	München, Nationalsozialistischer Untergrund, Prozess	yes
2	16	3	0.205	Berlin, CDU, Frauenquote	yes
3	15	3	0.242	Hugo Chavez, Nicolas Madura, Venezuela	yes
4	15	3	0.222	Euro, Polizei, Polizei	no
5	12	4	0.243	Polizei, Polizeipräsidium, Polizeipräsidium Mittelfranken, Rat	yes
6	11	5	0.244	Außenminister, John Forbes Kerry, Nordkorea, Tokio, Verhandlung	yes
7	11	6	0.237	Bundesärztekammer, Chef, Isar, Klinikum, München, Vorwurf	yes
8	10	4	0.346	Berlin, Deutschland, Jordanien, Syrien	yes
9	8	6	0.313	DFB,FC Bayern München, Halbfinale, Pokal, VfL Wolfsburg	yes
10	8	5	0.270	ARD, ZDF, ZDFinfo, ZDFkultur, ZDFneo	no

Table 3: Exemplary list of the top 10 topics of the 15th of April 2013 calculated using our algorithm. Its shows the rank, the number of articles belonging to the topic, the number of their common semantic items, and the similarity between the articles. Further, we try o generate labels for clusters using the (German) names of the SemItems. The last column relies on a manual, subjective rating of users if it is a topic or not.

our locations with coordinates retrieved from GeoNames. Unfortunately, the tool support for clustering RDF instances is rare. We only identified and successfully tested a RapidMiner plug-in called RMonto [9]. However, it is not feasible for our enterprise use case as the prototypical implementation is not fast enough and provides only a small set of algorithms. But as mentioned in Sect. 4.2, we could rely on the mining features supplied by Oracle. Currently, this step is an open issue.

4.4 Search and User Interface

Finally, to enable the access to the novel organized news stream we employ a search component ④. It accesses the storage layer to build an index for every proper characteristic, e. g., author, agency, headlines, SemItems according their type, or topics. Thus, it enables to delve into the media assets with an ordinary keyword or faceted search quite fast. Furthermore, we are able to create generic or specific queries, e. g., give me all assets comprising 'Lance Armstrong' and 'Oprah Winfrey' within the last 5 days, which all in all will assist the journalists in their work.

For evaluation purpose, we currently rely on the flexible, widget- and web-based user interface (UI) of our newsroom application, which allows for the common keyword-based search (Fig. 4). It comprises a free text search ① which constraints the result list with proper articles ②. Selecting a list item shows its content ③ and its recognized named entities ④. Furthermore, facet widgets present available persons, organizations, locations, and categories which could be used for filtering in ⑤ as well. Since we also calculate the semantic relatedness between articles, associated articles for the selected one are shown in ⑥. The user is able to adapt the composite user interface according his needs by integrating, removing, or re-ordering the widgets.

At this project stage, we are designing two more sophisticated UI which are addressing especially the promotion of topics and trends investigated in the backend. The first UI targets big, passive screens attached at the walls within a newsroom. Here, the editors should get a fast overview of the most important topics and their trend without being distracted from the current work. The second UI concept addresses editors investigation of topics, their trends and associated media assets at his workplace. Like in the first concept, the user should get a quick overview as well but should also be able to delve into the data. Therefore, maybe the metro map approach [11] is a suitable one.

5. INSIGHTS ABOUT THE PRE-PROCESSING

Through the prototypical implementation of the workflow we identified some obstacles particularly with the recognition and disambiguation of named entities (NE). Altogether, the complete pre-processing step (Fig. 3-1) was originally not part of our research project as we expected to simply include well-known tools or services in our architecture to retrieve the non-ambiguous German and English named entities. To identify the best for our purpose, we had to evaluate their NE recognition and disambiguation (NER, NED) capabilities. We especially concentrate on the most mature types person, organization, and location.

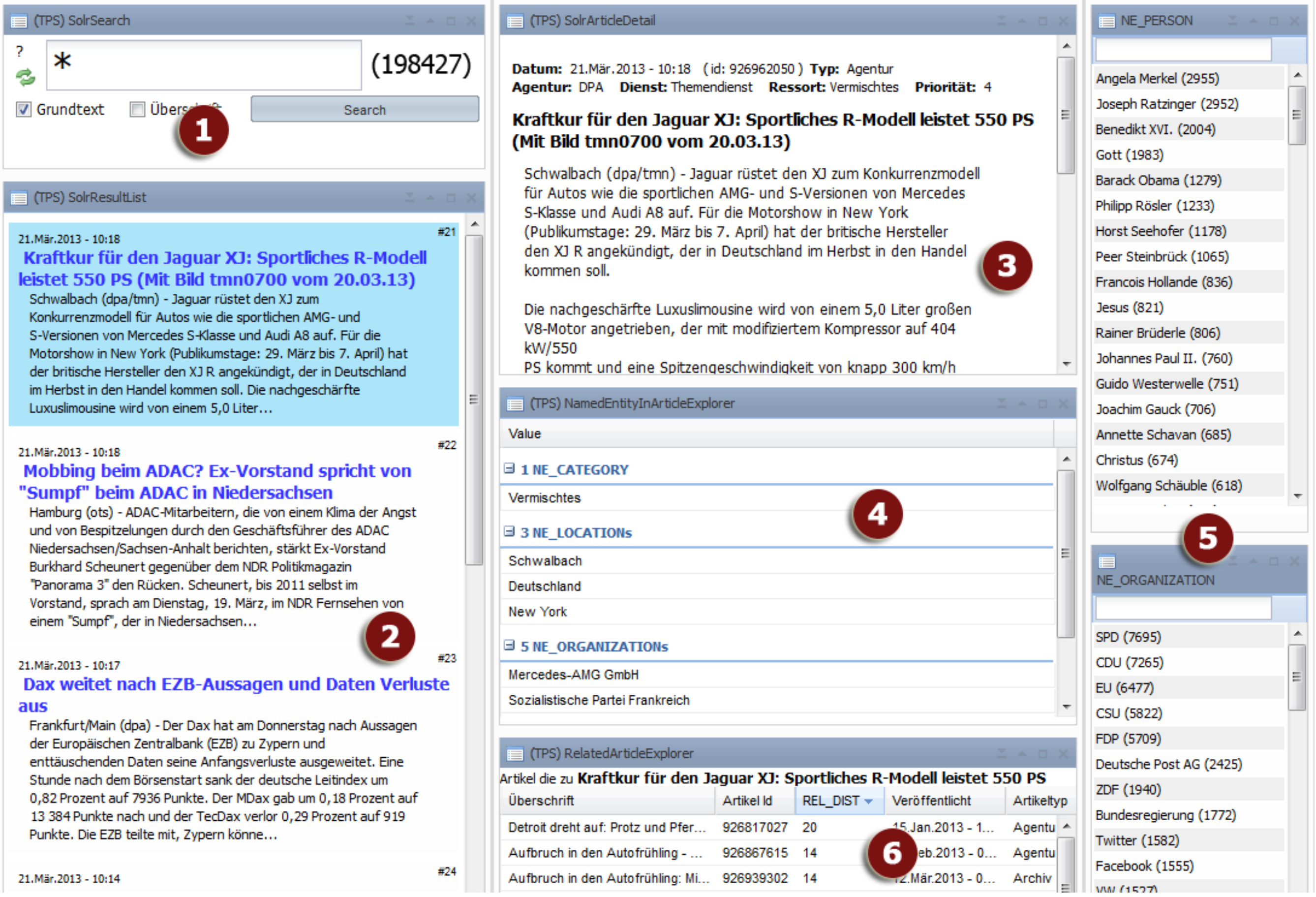
The NER for English news articles with common statistics-based tools performs like supposed, e. g., Stanford NER[15] reaches a usable F_1 score close to 90%. Unfortunately, the German NER judged with the CoNLL-2003 dataset[16] reaches only a F_1 score about 70% what not suffices our commercial use case and necessitates better statistical models. Due to this results we evaluated the wordlist-based approach, e. g., JRC-Names[17]. The tools provide an excellent precision of more than 90% but a bad recall of around 10% which is up to the quality of the word list. The method allows for a good configuration which named entities should be recognized but it is hard work to create the catalogs and keep them up-to-date. Further, we rely on instances from YAGO2 which comprises at most globally known entities and lack of local persons or organizations. It seems to be a good business model to offer formal thesauri for distinct domains or scopes.

As the mentioned approaches allow for extracting named entities they do not distinguish their specific type, e. g., *Paris* as city, name, or film. NED tools try to fill this gap and potentially offer a distinct URI [10]. But we identified

[15] `http://nlp.stanford.edu/software/`

[16] `http://www.cnts.ua.ac.be/conll2003/ner/`

[17] `http://ipsc.jrc.ec.europa.eu/index.php?id=42`



#21
21.Mär.2013 - 10:18
Kraftkur für den Jaguar XJ: Sportliches R-Modell leistet 550 PS (Mit Bild tmn0700 vom 20.03.13)
Schwalbach (dpa/tmn) - Jaguar rüstet den XJ zum Konkurrenzmodell für Autos wie die sportlichen AMG- und S-Versionen von Mercedes S-Klasse und Audi A8 auf. Für die Motorshow in New York (Publikumstage: 29. März bis 7. April) hat der britische Hersteller den XJ R angekündigt, der in Deutschland im Herbst in den Handel kommen soll. Die nachgeschärfte Luxuslimousine wird von einem 5,0 Liter...

#22
21.Mär.2013 - 10:18
Mobbing beim ADAC? Ex-Vorstand spricht von "Sumpf" beim ADAC in Niedersachsen
Hamburg (ots) - ADAC-Mitarbeitern, die von einem Klima der Angst und von Bespitzelungen durch den Geschäftsführer des ADAC Niedersachsen/Sachsen-Anhalt berichten, stärkt Ex-Vorstand Burkhard Scheunert gegenüber dem NDR Politikmagazin "Panorama 3" den Rücken. Scheunert, bis 2011 selbst im Vorstand, sprach am Dienstag, 19. März, im NDR Fernsehen von einem "Sumpf", der in Niedersachsen...

#23
21.Mär.2013 - 10:17
Dax weitet nach EZB-Aussagen und Daten Verluste aus
Frankfurt/Main (dpa) - Der Dax hat am Donnerstag nach Aussagen der Europäischen Zentralbank (EZB) zu Zypern und enttäuschenden Daten seine Anfangsverluste ausgeweitet. Eine Stunde nach dem Börsenstart sank der deutsche Leitindex um 0,82 Prozent auf 7936 Punkte. Der MDax gab um 0,18 Prozent auf 13 384 Punkte nach und der TecDax verlor 0,29 Prozent auf 919 Punkte. Die EZB teilte mit, Zypern könne...

#24
21.Mär.2013 - 10:14

Datum: 21.Mär.2013 - 10:18 (id: 926962050) **Typ:** Agentur
Agentur: DPA **Dienst:** Themendienst **Ressort:** Vermischtes **Priorität:** 4

Kraftkur für den Jaguar XJ: Sportliches R-Modell leistet 550 PS (Mit Bild tmn0700 vom 20.03.13)

Schwalbach (dpa/tmn) - Jaguar rüstet den XJ zum Konkurrenzmodell für Autos wie die sportlichen AMG- und S-Versionen von Mercedes S-Klasse und Audi A8 auf. Für die Motorshow in New York (Publikumstage: 29. März bis 7. April) hat der britische Hersteller den XJ R angekündigt, der in Deutschland im Herbst in den Handel kommen soll.

Die nachgeschärfte Luxuslimousine wird von einem 5,0 Liter großen V8-Motor angetrieben, der mit modifiziertem Kompressor auf 404
kW/550
PS kommt und eine Spitzengeschwindigkeit von knapp 300 km/h

Value

⊟ 1 NE_CATEGORY
Vermischtes

⊟ 3 NE_LOCATIONs
Schwalbach
Deutschland
New York

⊟ 5 NE_ORGANIZATIONs
Mercedes-AMG GmbH
Sozialistische Partei Frankreich

NE_PERSON
Angela Merkel (2955)
Joseph Ratzinger (2952)
Benedikt XVI. (2004)
Gott (1983)
Barack Obama (1279)
Philipp Rösler (1233)
Horst Seehofer (1178)
Peer Steinbrück (1065)
Francois Hollande (836)
Jesus (821)
Rainer Brüderle (806)
Johannes Paul II. (760)
Guido Westerwelle (751)
Joachim Gauck (706)
Annette Schavan (685)
Christus (674)
Wolfgang Schäuble (618)

NE_ORGANIZATION
SPD (7695)
CDU (7265)
EU (6477)
CSU (5822)
FDP (5709)
Deutsche Post AG (2425)
ZDF (1940)
Bundesregierung (1772)
Twitter (1582)
Facebook (1555)
VW (1527)

Artikel die zu **Kraftkur für den Jaguar XJ: Sportliches R-Modell leistet 550 PS**

Überschrift	Artikel Id	REL_DIST	Veröffentlicht	Artikeltyp
Detroit dreht auf: Protz und Pfer...	926817027	20	15.Jan.2013 - 1...	Agentu
Aufbruch in den Autofrühling - ...	926867615	14	...eb.2013 - 0...	Agentu
Aufbruch in den Autofrühling: Mi...	926939302	14	...2.Mär.2013 - 0...	Archiv

Figure 4: Screenshot of the widget-based newsdesk user interface

mainly two issues. First, with the exception of Alchemy[18] the APIs focus on English texts. Thus, the F_1 score for German news articles are lower than 40%. Hence, we strive for developing our own disambiguation algorithms (Sect. 4.1). Second, the NED functionality is mostly provided as an API, which is in a commercial use case often not feasible because of missing rights to send data to a third party. Only DBpedia spotlight [8], which is a research prototype, is freely available and could be installed locally.

Finally, a big issue is the absence of proper datasets to evaluate the NER and NED approaches. The manual creation of a set is a complex and time-consuming task. On working with CoNLL-2003 we especially identified that the following characteristics should be targeted: accessibility, currentness, domain, multilingualism, size, and the uniqueness of entities.

Accessibility The development and management of test corpora is costly, thus, fees for their usage are demanded. A further problem are the application of proprietary formats. With the move towards open data[19] research founding should also address such benchmark datasets. The *NLP Interchange Format* (NIF) approach[20] seems to be a step in the right direction.

Currentness Many datasets are kind of outdated as they rely on data which is older then 10 years. For instance in the news domain many new persons and organizations arise which are not reflected. Further, the writing style of journalists changes over decades what may have impact on the measured precision and recall.

Domain Every domain has its own concepts and, hence, requires specific datasets. It is especially difficult to create test sets for broad fields like the news domain.

Multilingualism Many of the datasets are English-only what may be one reason for the good quality and availability of English NER tools. But there is a growing interest for NER and NED also in other languages [3].

Size Although the size of the annotated text corpus matters for instance if a statistical model needs to be trained. The more text are available the better the results.

Uniqueness of entities Maybe one of the biggest problems to evaluate NED approaches are the missing links to knowledge bases, e. g., DBpedia. It es a cumbersome task to add them to existing datasets like CoNLL-2003.

[18] http://www.alchemyapi.com/
[19] http://5stardata.info/
[20] http://nlp2rdf.org/about

6. CONCLUSIONS AND FUTURE WORK

In this paper we gave a brief overview of our Topic/S vision and workflow which allows for the automatic retrieval of non-ambiguous named entities and categories from media assets. Furthermore, it allows for clustering the semantic items to identify topics, their relation among each other, and their trends. In future, these recommendations will enhance and speed-up journalists work as they will not miss any subject and its relevant articles, photos, or videos. We also identified some issues with regard to the relatively "old" discipline of named entity recognition which challenged us during our work. In particular, the tools do mostly neglect non-English customers and users. It is a time-consuming and costly task to achieve a high recognition quality which is required for a commercial use case.

Our current efforts are targeting the enhancement of our topic detection but also the disambiguation of named entities using our own semantic model as well as external resources. As mentioned in Sect. 4.4, we are working on user interfaces tailored to the needs of editors to identify and browse the topics. The last project milestone is to evaluate our approach in the daily business of selected customers.

7. ACKNOWLEDGMENTS

Work on this paper is funded within the Topic/S project by the European Regional Development Fund / Free State of Saxony, contract no. 99457/2677.

8. REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: Fighting Fire with Information from Social Web Stream. In *International Conference on Hypertext and Social Media, Milwaukee, USA*. ACM, 2012.

[2] C. Best, E. van der Goot, K. Blackler, T. Garcia, and D. Horby. Europe media monitor - system description. Technical Report EUR 22173 EN, European Commission, 2005.

[3] M. Ehrmann, M. Turchi, and R. Steinberger. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of Recent Advances in Natural Language Processing*, pages 118–124, 2011.

[4] T. Groza, S. Handschuh, and K. Moeller. The NEPOMUK Project - on the way to the social semantic desktop. Technical report, Digital Enterprise Research Institute (DERI), 2007.

[5] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.

[6] J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 229–232, New York, NY, USA, 2011. ACM.

[7] E. Klemm. *Das Problem der Distanzbindungen in der hierarchischen Clusteranalyse*. Lang, Frankfurt am Main [u.a.], 1995.

[8] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.

[9] J. Potoniec and A. ÅĄawrynowicz. Rmonto: Ontological extension to rapidminer. In *10th International Semantic Web Conference (ISWC2011)*, 2011.

[10] G. Rizzo and R. Troncy. Nerd: Evaluating named entity recognition tools in the web of data. In *Workshop on Web Scale Knowledge Extraction, ISWC*, 2011.

[11] D. Shahaf, C. Guestrin, and E. Horvitz. Trains of thought: generating information maps. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 899–908, New York, NY, USA, 2012. ACM.

[12] M. Voigt, A. Mitschick, and J. Schulz. Yet another triple store benchmark? practical experiences with real-world data. In *2nd International Workshop on Semantic Digital Archives (SDA2012)*, 2012.