Report from Dagstuhl Seminar 15201

Cross-Lingual Cross-Media Content Linking: Annotations and Joint Representations

Edited by

Alexander G. Hauptmann [1], James Hodson [2], Juanzi Li [3], Nicu Sebe [4], and Achim Rettinger [5]

1 Carnegie Mellon University, US, alex@cs.cmu.edu
2 Bloomberg New York, US, jhodson2@bloomberg.net
3 Tsinghua University Beijing, CN, lijuanzi@tsinghua.edu.cn
4 University of Trento, IT, sebe@disi.unitn.it
5 KIT Karlsruher Institut für Technologie, DE, rettinger@kit.edu

Abstract

Dagstuhl Seminar 15201 was conducted on Cross-Lingual Cross-Media Content Linking: Annotations and Joint Representations. Participants from around the world participated in the seminar and presented state-of-the-art and ongoing research related to the seminar topic. An executive summary of the seminar, abstracts of the talks from participants and working group discussions are presented in the forthcoming sections.

Seminar May 10 13, 2015 http://www.dagstuhl.de/15201
1998 ACM Subject Classi cation I.2.7 Natural Language Processing, I.2.10 Vision and Scene Understanding, H.3.3 Information Search and Retrieval, I.2.4 Knowledge Representation Formalisms and Methods
Keywords and phrases Cross-lingual, Cross-media, Cross-modal, Natural language processing, Computer vision, Multimedia, Knowledge representation, Machine learning, Information extraction, Information retrieval
Digital Object Identi er 10.4230/DagRep.5.5.43
Edited in cooperation with Aditya Mogadala

1 Executive Summary

Alexander G. Hauptmann
James Hodson
Juanzi Li
Nicu Sebe
Achim Rettinger

Di erent types of content belonging to multiple modalities (text, audio, video) and languages are generated from various sources. These sources either broadcast information on channels like TV and News or allow collaboration in social media forums. Often multiple sources are consumed in parallel. For example, users watching TV tweeting their opinions about a show. This kind of consumption throw new challenges and require innovation in the approaches to enhance content search and recommendations.

Cross-Lingual Cross-Media Content Linking: Annotations and Joint Representations, Dagstuhl Reports , Vol. 5, Issue 5, pp. 43 56
Editors: Alexander G. Hauptmann, James Hodson, Juanzi Li, Nicu Sebe, and Achim Rettinger
Dagstuhl Reports
Schloss Dagstuhl Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Currently, most of search and content based recommendations are limited to monolingual text. To nd semantic similar content across di erent languages and modalities, considerable research contributions are required from various computer science communities working on natural language processing, computer vision and knowledge representation. Despite success in individual research areas, cross-lingual or cross-media content retrieval has remained an unsolved research issue.

To tackle this research challenge, a common platform is provided in this seminar for researchers working on di erent disciplines to collaborate and identify approaches to nd similar content across languages and modalities. After the group discussions between seminar participants, two possible solutions are taken into consideration:

1. Building a joint space from heterogeneous data generated from di erent modalities to generate missing or to retrieve modalities. This is achieved through aligned media collections (like parallel text corpora). Now to nd cross-media cross-lingual relatedness of the content mapped to a joint latent space, similarity measures can be used.
2. Another way is to build a shared conceptual space using knowledge bases(KB) like DBpedia etc for semantic annotation of concepts or events shared across modalities and languages. Entities are expressed in any channel, media type or language cam be mapped to a concept space in KB. Identifying a commonality between annotations can be used to nd cross-media cross-lingual relatedness.

Thus, implementing these solutions require a joint e ort across research disciplines to relate the representations and to use them for linking languages and modalities. This seminar also aimed to build datasets that can be used as standard test bed and benchmark for cross-lingual cross-media content linking. Also, seminar was very well received by all participants. There was a common agreement that the areas of text, vision and knowledge graph should work more closely together and that each discipline would bene t from the other. The participants agreed to continue to work on two cross-modal challenges and discuss progress and future steps in a follow-up meeting in September at Berlin .

2 Table of Contents

Executive Summary
Alexander G. Hauptmann, James Hodson, Juanzi Li, Nicu Sebe, and Achim Rettinger 43

Overview of Talks

NLU for Colloquial Text and Speech-to-text
Xavier Carreras 46

Data Analytics over Mulitple Content Types
John Davies 47

Cross-Domain Cue Switching
Tiansi Dong 47

Cross-Lingual Document Similarity and Event Tracking
Blaz Fortuna 48

NELL as a Knowledge Graph building tool
Estevam R. Hruschka 48

Multi Lingual Knowledge Graph
Juanzi Li 49

Extracting aggregated knowledge from cross-lingual news
Dunja Mladenic 49

Multimodal Learning
Aditya Mogadala 50

Bloomberg Named Entity Disambiguation
Stefano Paci co 50

Machine Learning, Image Annotation and Computer Vision
Alan Smeaton 51

Relational Machine Learning for Knowledge Graphs
Volker Tresp 51

Automatic extraction of ontology lexica in multiple languages
Christina Unger 52

Learning Knowledge Graphs from Images and Text
Lexing Xie 52

Working Groups

Working Group II: State-of-the-art Text and Knowledge Graphs
Estevam R. Hruschka 53

Working Group III: Visual Information and Knowledge Graphs
Dubravko Culibrk 53

Working Group IV: Representation Learning
Aditya Mogadala 54

Open Problems 55

Participants 56

3 Overview of Talks

3.1 NLU for Colloquial Text and Speech-to-text

Xavier Carreras (Xerox Research Centre Europe Grenoble, FR)

State-of-the-art approaches to Natural Language Understanding (NLU) are based on supervised statistical techniques, and thus rely on the availability of treebanks, i.e. textual collections annotated with linguistic structure. Most available treebanks today annotate newswire articles, which are characterized by being edited text written by professionals. In contrast, many NLU applications deal with text generated by non-professional writers, which in most cases is produced spontaneously in conversations. This is the case for most of web data, emails, dialogue systems, and social media chatter. In these type of textual data, it is common to nd spelling mistakes and ungrammatical constructs. In addition, the distribution of topics and words will in most cases di er signi cantly from newswire data. All these facts pose di cult challenges for NLU on colloquial text. Similarly, in applications dealing with speech, the automatic conversion from speech to text results in noisy textual data which signi cantly di ers from the mistake-free text we nd in treebanks.

In this talk I will review recent work in the state-of-the-art for NLU on colloquial text and speech data. I will describe work in three directions. The rst is applying domain adaptation techniques, where I will review the main conclusions of the recent Parsing the Web challenge [1], and I will describe the architecture of the top-performing system (Le Roux et al 2012).

Then I will describe work for linguistic analysis of Twitter by Gimpel et al. [2], Owoputi et al. [3], and Kong et al. [4]. In these approaches, tweets are seen so dramatically di erent than standard newswire data that authors choose to rede ne the classic annotation standards and reannotate data. In other words, authors start a new task from scratch, attempting to come up with linguistic annotations that re ect the nature of tweets, as opposed to trying to describe tweets as if they were news articles. Some challenges here are how to deal with spontaneous variations of tokens, and how to take advantage of the large and rich resources we have for standard domains.

Finally I will describe some approaches for Spoken Language Understanding. To cope with the noisy input, one approach is to directly represent the confusion word lattice with various features that capture the uncertainty.

Overall, these techniques show a degree of improvements. But the problem of NLU on colloquial text and speech-to-text remains a challenging, di cult and important open problem.

References

1 Petrov, S., and McDonald, R.: Overview of the 2012 shared task on parsing the web. In Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL) (2012)
2 Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., and Smith, N. A. Part-of-speech tagging for twitter: Annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.Association for Computational Linguistics. (2011)

3 Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics. (2013)
4 Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. A dependency parser for tweets. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2014)

3.2 Data Analytics over Mulitple Content Types

John Davies (BT Research Ipswich, GB)

We present 3 use cases of the use of data analytics over multiple content types. In the rst case, structured data and unstructured data (text documents) relating to an organization's sales performance are analyzed. Named entity recognition is used to identify people, products, companies, locations and key-phrases in the textual data and associated with an ontology using semantic annotation. The structured data is held in a traditional RDB and a dashboard is developed allowing queries over both data types allowing the data to be analyzed for improved sales management information in ways previously only possible via a time-consuming manual process. In the second use case, we combine sensor data with social media data. We exemplify in the transport domain, where tweets about tra c are ltered and their location is extracted automatically from the tweet content. This can then be combined with roadside sensor data giving a combined view of both numeric values such as tra c speed and density and Twitter users observations about the same road segment. This is useful for highways authorities where valuable information relating to tra c incidents in often to be found in social media content. Finally, we discuss the value of combining video content with social media data and speci cally of aligning events described or shown in both media. Here the example is sport, where incidents in sports games could be detected by video recognition (and/or textual analysis of subtitles)and linked to social media users' reactions to the same event. Media companies can thereby gauge reaction to certain events and better understand their customers.

3.3 Cross-Domain Cue Switching

Tiansi Dong (Universität Bonn, DE)

Descriptions of the same entity may capture di erent meaning aspects, if we choose di erent media. Cue, as used mainly in psychology, refers to any piece of information between descriptions and meaning aspects. To transform a description from the source media into a form in the target media, we shall rst retrieve the meaning aspect of the description in the source domain, transform it into the meaning aspect of the target media, and deliver descriptions of the target media. Three examples in natural language translation are presented. The rst example is to translate white as snow into the native language of Benin, where there is no word for snow ; the second example is to translate you are my heart into

Indonesian , where hearts are regarded less important than livers; the third example is to translate the description the western table into the table on my left , which can only be achieved through a spatial transformation. Such spatial transform can be achieved by understanding orientation relations as distance comparison relations. We further show how orientation relations shall be understood as distance comparison relations in general, and how distance comparison and distance relations can be de ned in the connection relation. As spatial domain is the rst domain human babies encounter and understand, this domain is used as the reference domain for the cognition of other domains. Thus, spatial domain is the base domain for cross-domain cue switching. Our current research work on German-Chinese cue-switching translation is outlined. Cues used in other domains are listed.

3.4 Cross-Lingual Document Similarity and Event Tracking

Blaz Fortuna (Ghent University, BE)

Joint work of Rupnik, Jan; Muhic, Andrej; Leban, Gregor; Skraba, Primoz; Fortuna, Blaz; Grobelnik, Marko;
URL http://xling.ijs.si/

In this work, we address the problem of tracking and events in a large multilingual stream. We consider a particular aspect of this problem, namely how to link collections of articles in di erent languages which refer to the same event.

Given a multi-lingual stream and clusters of articles from each language, we propose a method for cross-lingual document similarity based on Wikipedia, which enables us to compute the similarity of any two articles regardless of language. The approach learns an representations of documents which were valid over multiple languages. The representations could be interpreted as multi-lingual topics, which were then used as proxies to compute cross-lingual similarities between documents. To learn the representations, we use Wikipedia as a training corpus. Signi cantly, we do not only consider the major or hub languages such as English, German, French, etc. which have signi cant overlap in article coverage, but also smaller languages (in terms of number of Wikipedia articles) such as Slovenian and Hindi, which may have a negligible overlap. The proposed method can scale to 100 languages and can match articles from languages with little or no direct overlap in the training data.

3.5 NELL as a Knowledge Graph building tool

Estevam R. Hruschka (University of Sao Carlos, BR)

Joint work of Mitchell, Tom M.; Cohen, William W.; Hruschka, Estevam R.
Main reference T. M. Mitchell, W. W. Cohen, E. R. Hruschka Jr., P. P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi Mishra, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. A. Platanios, A. Ritter, M. Samadi, B. Settles, R. C. Wang, D. T. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, J. Welling, Never-Ending Learning, in Proc. of the 29th AAAI Conf. on Arti cial Intelligence (AAAI'15), pp.>2302 2310, AAAI Press, 2015.
URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10049

Never-Ending Language Learner (NELL) is a computer system that runs 24/7, forever, learning to read the web. The system is designed to perform two basic tasks: i) extract (read) more facts from the web, and integrate these into its growing knowledge base of beliefs; and

ii) learn to read better than yesterday, enabling it to go back to the text it read yesterday, and today extract more facts, more accurately. This system has been running 24 hours/day for over four years now. The result so far is a collection of 90 million interconnected beliefs (e.g., servedWith(co ee, applePie), isA(applePie, bakedGood)), that NELL is considering at di erent levels of con dence, along with hundreds of thousands of learned phrasings, morphological features, and web page structures that NELL uses to extract beliefs from the web.

3.6 Multi Lingual Knowledge Graph

Juanzi Li (Tsinghua University Beijing, CN)

Multilingual knowledge graph is the graph of entities and relationships in di erent languages. Multilingual knowledge graph, recognized as the bridges for information understanding across multiple languages, can enhance the linked data internationalization and globalization of knowledge sharing among di erent languages on the Web and, facilitate the cross-lingual language processing such as cross-lingual information retrieval, machine translation and question answering etc. In this talk, we summarize the existing monolingual and multilingual knowledge graphs, and the state of the art technologies including multilingual knowledge linking, multilingual knowledge building and cross lingual knowledge extraction. The talk is concluded with some identi ed challenging problems such us multilingual knowledge representation learning, multimedia and multilingual knowledge linking based on it.

3.7 Extracting aggregated knowledge from cross-lingual news

Dunja Mladenic (Jozef Stefan Institute Ljubljana, SI)

Joint work of Mladenic, Dunja; Marko Grobelnik; Blaz Fortuna; Gregor Leban; Blaz Noak; Jan Rupnik; Mitja Trampus; Andrej Muhic

Cross-lingual news analysis in general requires handling large amount of textual data across di erent languages. We propose combining Machine Learning and Natural Language Processing methods to enable extracting aggregated knowledge from cross-lingual news. In particular, we propose several lines of development involving research and development of prototype systems for news annotation in real-time, mining event patterns, identifying events across languages, detect diversity of reporting along several dimensions, rich exploratory visualizations of news events, interoperable data export.

We demonstrate the functioning on operational news monitoring and extracting knowledge that involves a large number of data streams in multiple languages. The following are publicly available related systems: Event Registry[1] for event detection and topic tracking;

[1] http://eventregistry.org//

DiversiNews[2] for news diversity explorer; NewsFeed[3] for news and social media crawler; Enrycher[4] for language and semantic annotation; XLing[5] for cross-lingual document linking and categorization.

3.8 Multimodal Learning

Aditya Mogadala (KIT – Karlsruher Institut für Technologie, DE)

The growth of multimedia content on the web raise diverse challenges. Over the decades various approaches are designed to support search, recommendations, analytics and advertising based on the textual content. But now due to the overwhelming availability of multimedia content require update in technologies to leverage multimedia information. Recent advancements made in machine learning to foster continuous representations of text and e ectual object detection in videos and images provide new opportunities. In this aspect, leveraging data generated from videos, images and text to support various applications by nding cross-modal semantic similarity. In particular, to compare semantically similar content generated across media by jointly modeling two di erent modalities.

Modeling one or more modalities together can be helpful to generate missing modalities and retrieve cross-modal content. Results are overwhelming when used to jointly model images or videos along with captions using deep learning approaches like Recurrent neural networks and multimodal log-bilinear models. It also pave the path to extend textual information to multiple languages for supporting the growth of polylingual content on the web.

3.9 Bloomberg Named Entity Disambiguation

Stefano Paci co (Bloomberg – New York, US)

Joint work of Paci co, Stefano; Bradesko, Luka; Starc, Janez
Main reference L. Bradesko, J. Starc, S. Paci co, Stefano, Isaac Bloomberg Meets Michael Bloomberg: Better Entity Disambiguation for the News, in Proc. of 24th Int'l World Wide Web Conference (WWW'15) Companion Volume, pp. 631 635, ACM, 2015.
URL http://dx.doi.org/10.1145/2740908.2741711

In this talk we present BNED, the Named Entity Disambiguation system developed and used at Bloomberg. In particular, we illustrate how we built a system that do not require the use of Wikipedia as a knowledge base or training corpus. We also present how we built features for disambiguation algorithms signi cative for the Bloomberg News corpus, and show results of both single-entity and joint-entity disambiguation into the Bloomberg proprietary knowledge base of people and companies.

[2] http://aidemo.ijs.si/diversinews//
[3] http://newsfeed.ijs.si//
[4] http://enrycher.ijs.si//
[5] http://xling.ijs.si//

3.10 Machine Learning, Image Annotation and Computer Vision

Alan Smeaton (Dublin City University, IE)

This presentation covered an overview of the stat-of-the-art in automatic detection of semantic concepts from visual media. The presentation started with an overview of the challenges in captioning or tagging or annotating visual media and then described how we use low-level image features colours, textures, shapes, SIFT/SURF features to index images so we can support look-alike visual similarity searching. This is useful in some applications but doesn't directly address the problem of describing an image's contents. We then moved on to present how the multimedia indexing eld uses o -the-shelf machine learning to build classi ers, usually one at a time, and how the performance of these has evolved and improved over the last decade in the TRECVid benchmarking. We then looked at most recent work on image captioning using deep learning from groups in Stanford and in Google, as reported in the NYT, and nally ended the presentation with a showcase of building classi ers in real time, during a search, the advantages of this approach being that we don't have to know what people might want to search for in advance.

3.11 Relational Machine Learning for Knowledge Graphs

Volker Tresp (Siemens AG München, DE)

Joint work of Tresp, Volker; Maximilian Nickel; Denis Krompass; Xueyan Jiang
Main reference M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, A Review of Relational Machine Learning for Knowledge Graphs: From Multi-Relational Link Prediction to Automated Knowledge Graph Construction, to appear as invited paper in the Proceedings of the IEEE; pre-print available as arXiv:1503.00759v3 [stat.ML].
URL http://arxiv.org/abs/1503.00759v3

Most successful applications of statistical machine learning focus on response learning or signal-reaction learning where an output is produced as a direct response to an input. An important feature is a quick response time, the basis for, e.g., real-time ad-placement on the Web, real-time address reading in postal automation, or a fast reaction to threats for a biological being. One might argue that knowledge about speci c world entities and their relationships is necessary if the complexity of an agent's world increases, for example if an agent needs to function in a complex social community. As one is quite aware in the Semantic Web community, a natural representation of knowledge about entities and their relationships is a directed labeled graph where nodes represent entities and where a labeled link stands for a true fact. A number of successful graph-based knowledge representations, such as DBpedia, YAGO, or the Google Knowledge Graph, have recently been developed and are the basis of applications ranging from the support of search to the realization of question answering systems. Statistical machine learning can play an important role in knowledge graphs as well. By exploiting statistical relational patterns one can predict the likelihood of new facts, nd entity clusters and determine if two entities refer to the same real world object. Furthermore, one can analyze new entities and map them to existing entities (recognition) and predict likely relations for the new entity. These learning tasks can elegantly be approached by rst transforming the knowledge graph into a 3-way tensor where two of the modes represent the

entities in the domain and the third mode represents the relation type. Generalization is achieved by tensor factorization using, e.g., the RESCAL approach. A particular feature of RESCAL is that it exhibits collective learning where information can propagate in the knowledge graph to support a learning task. In the presentation the RESCAL approach will be introduced and applications of RESCAL to di erent learning and decision tasks will be presented.

3.12 Automatic extraction of ontology lexica in multiple languages

Christina Unger (Universität Bielefeld, DE)

Many applications that need to mediate between natural language and Semantic Web data, such as question answering and verbalization of ontologies or RDF datasets, require knowledge about how elements of the vocabulary are expressed in natural language. Moreover, in case a system is supposed to be multilingual, this knowledge is needed in multiple languages. In this talk, I present a model for capturing such lexical knowledge as well as a recent approach to automatically acquiring it, and outline the main limitations this approach still faces.

3.13 Learning Knowledge Graphs from Images and Text

Lexing Xie (Australian National University Canberra, AU)

Main reference L. Xie, H. Wang, Learning Knowledge Bases for Text and Multimedia , ACM Multimedia 2014 Tutorial, 2014.
URL http://users.cecs.anu.edu.au/~xlx/proj/knowledge_mm14.html

Knowledge acquisition, representation, and reasoning have been one of the long-standing challenges in arti cial intelligence and related application areas. Only in the past few years, massive amounts of structured and semi-structured data that directly or indirectly encode human knowledge became widely available, turning the knowledge representation problems into a computational grand challenge with feasible solutions in sight. The research and development on knowledge bases is becoming a lively fusion area among web information extraction, machine learning, databases and information retrieval, with knowledge over images and multimedia emerging as another new frontier of representation and acquisition. This tutorial aims to present a gentle overview of knowledge bases on text and multimedia, including representation, acquisition, and inference. I present a brief survey of work on learning words, entities and their relations from images and their accompanying words.

4 Working Groups

4.1 Working Group II: State-of-the-art Text and Knowledge Graphs

Estevam R. Hruschka (University of Sao Carlos, BR)

One subgroup was focused on discussing and motivating all the participants to discuss the state-of-the-art bridging Text and Knowledge Graphs, and how Cross-Lingual, as well as Cross-Media can be explored to help in de ning joint Representations. The discussions led to some very interesting issues, and most of those are based on the fact that it is very di cult to de ne an optimum representation (ontology) describing an ideal taxonomy that would allow:

(i) Using the current state-of-the-art to generate more useful results (more useful knowledge graphs) for real problems applications;
(ii) Identi cation of crucial research challenges and key points that should drive research e orts in the near future.

Based on the aforementioned items, some more discussions were motivated and the group tried to formulate possible tangible ways to cope with the representation problems. The main goal was to identify concrete ways to achieve better results from possible future collaborations and follow-up actions that might start after this seminar.

The summary of the last discussions is:

1. We can put together e orts already being done on information extraction (from di erent languages and also images) and knowledge graph building. One concrete actions include coupling Chinese information extraction to NELL system.
2. We can de ne a common application domain in which we could apply the results of the collaboration proposed in item 1 (above). One concrete example would be using information extraction and knowledge bases to identify events in sports games matches (i.e. penalty in a soccer match) and associate that event with social media (i.e. Twitter) real time discussions.
3. The open questions (that should motivate research e orts) are:
   (i) How to evaluate the obtained results?
   (ii) How to have a robust representation for di erent domains?
   (iii) How to cope with temporal-spatial scope?
   (iv) How to put together deeper ontologies (CyC) with shallow representations such a s Knowledge graphs?

4.2 Working Group III: Visual Information and Knowledge Graphs

Dubravko Culibrk (University of Trento, IT)

Joint work of Witbrock, Michael;Grobelnik, Marko;Hodson, James;Paci co, Stefano;Novak, Blaz

The working group discussions started with the examination whether the Knowledge Graphs (KGs) are in fact necessary and useful to aid the vision tasks? Eventually a consensus was reached that the existing Knowledge Bases (KBs), i.e. KGs do not have enough coverage of to help disambiguate and aid the vision tasks. Therefore the discussion from that point on

focused on how to achieve the required extension of the KBs and the density of KGs required to make them useful in scenarios of relevance to vision. To achieve this practically we thought we could limit ourselves to some speci c tasks (such as understanding tabletops or meals) and extend existing KBs automatically. We would like to infer the world model for this domain, which could then be extended to other domains. We would like to model entities, relations, scripts. After the discussions with all the participants we extended our scenario to understanding everything that happens in a kitchen. The nal discussions revolved around how we could go about setting up a challenge that would help build a dense model of the world of kitchen. We would need a lot of video data, which is publicly available. To help the participants we would like to have the automatic speech recognition transcripts for the video data in the dataset. The tasks would need to be de ned to favour the teams that are able to e ectively use the multimodal data and require the inference of the model. The nal conclusion of this workgroup is that we would like to organise a challenge along the lines of has been discussed and pursue options to get EU funding to create and run the challenge over the next few years.

4.3 Working Group IV: Representation Learning

Aditya Mogadala (KIT Karlsruher Institut für Technologie, DE)

Joint work of Carreras, Xavier; Smeaton, Alan; Sebe, Nicu; Chong-Wah, Ngo; Thalhammer Andreas; Rettinger Achim

The working group discussions started with possible scenarios where multimedia can be leveraged with textual information and vice versa. Initially, an idea of using Google[6] image search was pitched in to disambiguate textual queries of multimedia search. An idea of building on the y image classi er by crawling rst 1000 images from Google to disambiguate retrieved results. This paved the path in a new direction to understand the images which lack objects and are very abstract. For example, how can a computer understand images that means emptiness. This drove the discussions to explore external information provided in the form of text or structured knowledge. Possibilities of using structured knowledge was inspected to identify relationships between objects detected in media content. Representation learning can be used in this scenario as:

- Good approach to predict missing modality.
- Application driven.
- Don't have to bother about features deep learning

Other ideas that are brainstormed use both multimedia and textual information for aligning religious pictures present in the museums along with ancient texts. Most of the picture galleries present in museums either lack descriptions or di cult to interpret their inherent depth. There are many ancient texts written that would have described something similar in the pictures. Aligning picture galleries to ancient texts manually can be tedious and cumbersome. Leveraging multimedia processing approaches with natural language understanding techniques can automate this process to an extent. Few other ideas that was discussed are:

1. Identifying a dominant person in a video where two people are debating on a topic.
2. Identifying a person and his role in the news domain.

[6] https://google.com

3. Predicting price from the product catalog.
4. Event identi cation about disasters

Further action was to come up with a benchmark dataset to solve any of these tasks and conduct challenge in future.

5 Open Problems

There are several open issues which needs to addressed. Few of them are listed below.

- Identifying if generative models work for cross-lingual and cross-media linking.
- What kind of approach that needs to be employed, if we do not have enough multilingual and multimedia training data.

Participants

- Xavier Carreras
Xerox Research Centre Europe
Grenoble, FR
- Dubravko Culibrk
University of Trento, IT
- John Davies
BT Research Ipswich, GB
- Tiansi Dong
Universität Bonn, DE
- Anastassia Fedyk
Harvard University, US
- Blaz Fortuna
Ghent University, BE
- Marko Grobelnik
Jozef Stefan Institute
Ljubljana, SI
- Alexander G. Hauptmann
Carnegie Mellon University, US
- James Hodson
Bloomberg New York, US
- Estevam R. Hruschka
University of São Carlos, BR
- Bea Knecht
Zattoo Zürich, CH
- Juanzi Li
Tsinghua University Beijing, CN
- Dunja Mladenic
Jozef Stefan Institute
Ljubljana, SI
- Aditya Mogadala
KIT Karlsruher Institut für Technologie, DE
- Chong-Wah Ngo
City University Hong Kong, HK
- Blaz Novak
Jozef Stefan Institute
Ljubljana, SI
- Stefano Paci co
Bloomberg New York, US
- Achim Rettinger
KIT Karlsruher Institut für Technologie, DE
- Evan Sandhaus
The New York Times, US
- Nicu Sebe
University of Trento, IT
- Alan Smeaton
Dublin City University, IE
- Rudi Studer
KIT Karlsruher Institut für Technologie, DE
- Jie Tang
Tsinghua University Beijing, CN
- Andreas Thalhammer
KIT Karlsruher Institut für Technologie, DE
- Eduardo Torres Schumann
VICO Leinfelden-Echterdingen, DE
- Volker Tresp
Siemens München, DE
- Christina Unger
Universität Bielefeld, DE
- Michael Witbrock
Cycorp Austin, US
- Lexing Xie
Australian National University
Canberra, AU
- Lei Zhang
KIT Karlsruher Institut für Technologie, DE