Bootstrapping an Online News Knowledge Base

Klesti Hoxha(✉), Artur Baxhaku, and Ilia Ninka

Department of Computer Science, Faculty of Natural Sciences,
University of Tirana, Blv. Zogu I, Nr. 20/1, 1001 Tirana, Albania
{klesti.hoxha,artur.baxhaku,ilia.ninka}@fshn.edu.al

Abstract. News retrieval systems facilitate the process of quickly learning about events or stories reported in various online news providers. The traditional approach involves clustering articles that report about the same event using bag-of-words or concept based similarity measures, and offering personalized recommendations using various user modeling approaches. Knowledge bases have been extensively used in the recent years for powering search engines on entity based searches. The success of this approach, demonstrated by a now de-facto way of searching and browsing offered by commercial search engines and mobile applications, has created the need to incorporate semantic capabilities to news retrieval systems. In this paper we present a proposal for creating a knowledge base of entities, events and facts reported in Albanian online news providers. We aim to provide a news stream processing pipeline based in generally available open source toolkits and state-of-the-art research works about event and fact oriented knowledge bases.

Keywords: News retrieval · Fact extraction · Event mining · Semantic news

1 Introduction

Knowledge bases (KB) are nowadays powering most of the commercial search engines[1]. They are mostly used to provide quick facts about people, organizations, sport teams and other entities related to the provided search queries. It has been shown that entity enriched search results provide a better user experience on the related systems [2].

Very often news articles are centered around particular entities: a politician's visit to a particular place, the result of a football match, a public figure speech related to a certain question, a terrorist attack on a city, annual cultural events, etc. Even though most of the available news retrieval systems offer related stories discovery, content grouping and even story development timeline visualizations, they still lack of knowledge discovery features generally available in modern search engines.

[1] Two concrete examples: Google Knowlege Graph https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html, and Microsoft Satori https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/.

A. Bozzon et al. (Eds.): ICWE 2016, LNCS 9671, pp. 501–506, 2016.
DOI: 10.1007/978-3-319-38791-8_37

Let's consider a visit of Angela Merkel to Albania. A news retrieval system that makes use of a knowledge base, would detect that a certain news article about this event is related to Angela Merkel and Albania, furthermore it would store the fact that it is about a *politician's visit* to a certain country. An advanced use of the knowledge base in question can recommend news articles about Angela Merkel's visits to other Balkan countries, or any country in general. It can also suggest articles about previous visits of her to Albania, previous Chanchellors of Germany visits to Albania, and other similar related articles. This entity centered personalization approach has been reported in some previous works [3,4,10].

In this paper we describe the requirements and initial steps for creating a news-centered knowledge base for Albanian written news articles published online. It will store facts about certain events reported in the news using a custom knowledge representation model. We define a system architecture that allows for different implementations of it. This architecture can be also used by news retrieval systems that deal with articles written in any other language, however considering the fact that the natural language processing landscape for Albanian is lacking many enabler components, we aim to facilitate extensive experimentation.

2 Related Work

Works reported in literature regarding news knowledge base creation focus on three main aspects: entity linking and disambiguation, knowledge graph representations (ontologies) for news events, and news processing pipelines for knowledge base population.

Entity linking, the process of relating named entities found in the text of the processed documents with existing entries in a knowledge base, deals with the need of entity disambiguation. It is the process of finding the correct entry in a KB for orthographically different mentions of an entity, or identifying missing entries [7]. It has been shown in different works that entity disambiguation for news articles is done considering the textual context of a named entity appearance and concept similarity graphs [5,8]. Skenduli and Biba [9] have demonstrated that named entities in Albanian can be accurately recognized using trained classifiers provided by Apache OpenNLP[2].

Our intent is to create a news knowledge base that contains information and facts about events or stories related to people or places. We initially plan to link the identified entities with existing entries of people and places in some publicly available knowledge bases like DBpedia[3] and Yago[4].

News processing pipelines for KB population reported in literature use a combination of tools for achieving this. Some of them introduce a service oriented architecture. Regarding event or facts extraction there are two main approaches: machine learning NLP techniques or rule-based and topic clustering methods.

[2] https://opennlp.apache.org/.

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

[4] https://www.mpi-inf.mpg.de/yago-naga/yago/.

An advanced multilingual news knowledge base is described by Rospocher et al. in [8]. They provide knowledge graphs of events reported in the news. It is created using a modular news processing pipeline with mostly custom build NLP tools for each involved language. Their approach processes a news collection all at once, not in an incremental manner.

XLike is another multilingual news processing pipeline [6]. It uses open source tools and generally available language corpora for implementing its NLP functionalities. News articles are clustered based on their topic, and a knowledge graph with facts and events is maintained. A similar architecture is described in [1], but lacking advanced NLP processing. It uses topic based clustering instead.

In [11] Zavarella et al. provide an example of a work that does not use a standard machine learning based NLP approach in its news processing pipeline. They describe a system that uses entity extraction grammars and semantic annotation through rule-based patterns. It is applied in crisis and security threat detection from news written in three Balkan languages.

3 Research Objectives

We aim to provide an initial setup of a news related knowledge base for news articles written in Albanian. Our main goal is to boost the user experience of news retrieval systems or news portals in general through advanced personalized news recommendation.

Due to the fact that the Albanian natural language processing landscape is still missing key components for creating advanced knowledge discovery systems, we can contribute in this regard as part of this work. This can be considered as another output of our research. In summary we have the following research objectives:

1. Propose a simple ontology for representing news events or facts.
2. Provide a software architecture for the news knowledge base that allows for extensive experimentation.
3. Contribute with corpora and tools for Albanian natural language processing.
4. Create an initial implementation of the proposed architecture using open source toolkits.

4 Methodology

We have started our work by developing a news aggregator for Albanian news using Scrapy[5]. In order to gain more context details (i.e. latest stories, important news) by the location of the page where the news is present, we do not use RSS feeds. News are stored in an intermediate representation using a NoSQL database (MongoDB[6]). For each article we also store extracted meta-data like publication

[5] http://scrapy.org/.

[6] https://www.mongodb.org/.

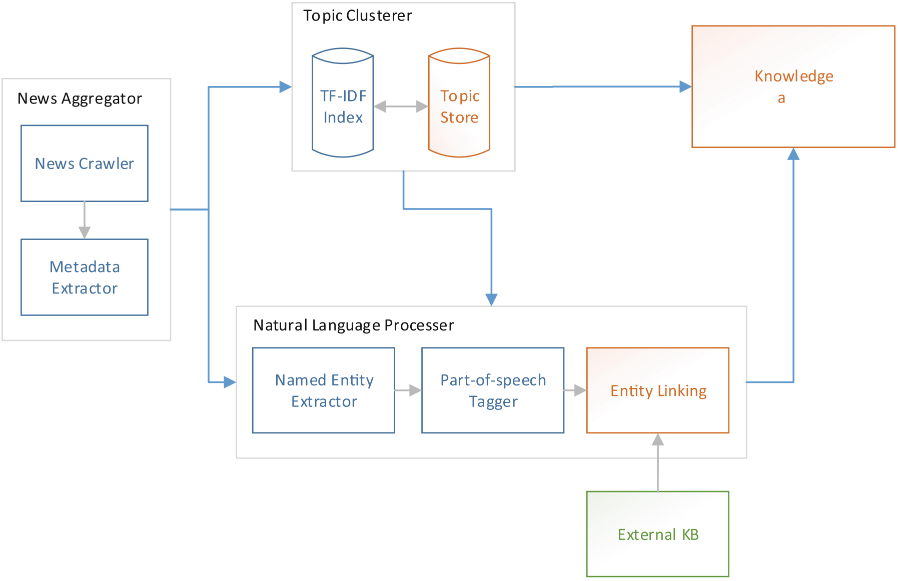

Fig. 1. Proposed architecture.

date, last update timestamp, author, extracted news category, number of comments, etc. We have also created a term-frequency index for the aggregated news using Apache Lucene[7]. This is used for clustering news articles about the same event using a term-frequency based similarity measure.

The proposed system architecture is shown in Fig. 1. We plan to use Apache OpenNLP for named entity recognition and part-of-speech tagging. Due to the lack of annotated corpora for this language, we are also creating them using the collected news as a corpus. Existing annotated corpora are also used for NLP processing in [6]. In order to allow experimentation with other NLP toolkits, we are using a custom annotation format that can be easily converted to the required format of the tool in question.

Because of the lack of quality annotated treebanks, we plan to skip machine learning techniques for semantic role labeling [6,8] and use a rule-based pattern matching approach similar to [11]. The set of events stored in the knowledge base will be limited in the initial stage. Table 1 shows a sample of the triples that will be created. For entity disambiguation [7] we plan to use the usage context with the help of the created term-frequency index. When linking to external knowledge bases we can also use location (for news about events happening in Albania) as a disambiguation feature. Entries in our knowledge base will also be linked to the source of the stored information, a single news or a topic cluster.

Considering that news article retrieval is a publication time sensitive task, the stream of news will be incrementally indexed and update the knowledge

[7] https://lucene.apache.org/.

Table 1. Sample triples included in the news KB.

Subject	Predicate	Object
Politician	spokeAbout	X
Politician	visited	Place
Parliament	approvedLaw	Law No.
Journalist	interviewed	Person
Artist	participatedIn	Concert
Concert	heldIn	Place
Accident	happenedIn	Place
Politician	met	Politician
SportMatch	endScore	X

base with new events or facts. The knowledge base will be accessible through a RESTFul API. This allows an easier integration to third-party systems like news search engines or news publication websites powering retrieval and personalized content offering.

5 Conclusions

In this work we describe our approach and initial steps on creating a knowledge base of events and stories reported in Albanian online news portals. We proposed an architecture of a news stream processing pipeline based on the current state-of-the-art solutions in this regard and implementable using various open source toolkits.

The initial plan is to offer access to the created knowledge base through a RESTFul API, however this can be extended also to the entity linking service of our system. This would allow the incorporation of advanced knowledge discovery features and facilitate personalized news recommendation to existing news search engines and publishing portals.

To the best of our knowledge, this is the first reported attempt to create a semantic knowledge base for documents written in Albanian. The datasets and annotated corpora created in this work will also contribute to the Albanian natural language processing landscape.

References

1. Amardeilh, F., Kraaij, W., Spitters, M., Versloot, C., Yurtsever, S.: Semi-automatic ontology maintenance in the virtuoso news monitoring system. In: 2013 European Intelligence and Security Informatics Conference (EISIC), pp. 135–138. IEEE (2013)

2. Arapakis, I., Leiva, L.A., Cambazoglu, B.B.: Know your onions: understanding the user experience with the knowledge module in web search. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1695–1698. ACM (2015)
3. Cadilhac, A., Chisholm, A., Hachey, B., Kharazmi, S.: Hugo: entity-based news search and summarisation. In: Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval, pp. 51–54. ACM (2015)
4. Hare, J., Newman, D., Peters, W., Greenwood, M., Eggink, J.: Semanticnews: Enriching publishing of news stories (2014)
5. Kuzey, E., Vreeken, J., Weikum, G.: A fresh look on knowledge bases: distilling named events from news. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 1689–1698. ACM (2014)
6. Padró, L., Agić, Ž., Carreras, X., Fortuna, B., Garcia-Cuesta, E., Li, Z., Štajner, T., Tadić, M.: Language processing infrastructure in the xlike project. In: Ninth International Conference on Language Resources and Evaluation (LREC 2014) (2014)
7. Rao, D., McNamee, P., Dredze, M.: Entity linking: finding extracted entities in a knowledge base. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, Multilingual Information Extraction and Summarization 11. Theory and Applications of Natural Language Processing, pp. 93–115. Springer, Heidelberg (2013)
8. Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T., Bogaard, T.: Building event-centric knowledge graphs from news. Sci. Serv. Agents World Wide Web, Web Seman. (2016)
9. Skenduli, M.P., Biba, M.: A named entity recognition approach for albanian. In: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1532–1537. IEEE (2013)
10. Tavakolifard, M., Gulla, J.A., Almeroth, K.C., Ingvaldesn, J.E., Nygreen, G., Berg, E.: Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 305–308. International World Wide Web Conferences Steering Committee (2013)
11. Zavarella, V., Kucuk, D., Tanev, H., Hürriyetoglu, A.: Event extraction for balkan languages. In: EACL 2014, p. 65 (2014)