Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 155-158 recibido 15-04-14 revisado 15-07-14 aceptado 15-07-14

NewsReader project

Proyecto NewsReader

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Begoña Altuna
Zuhaitz Beloki, Egoitz Laparra, Maddalen López de Lacalle
German Rigau, Aitor Soroa, Ruben Urizar
IXA NLP Group, University of the Basque Country UPV/EHU
Manuel Lardizabal Pasealekua 1, 20018 Donostia
german.rigau@ehu.es

Resumen: El proyecto europeo NewsReader desarrolla tecnología avanzada para procesar flujos contínuos de noticias diarias en 4 idiomas, extrayendo lo que pasó, cuándo, dónde y quién estuvo involucrado. NewsReader lee grandes cantidades de noticias procedentes de miles de fuentes. Se comparan los resultados a través de las fuentes para complementar la información y determinar en qué están de acuerdo. Además, se fusionak noticias actuales con noticias previas, creando una historia a largo plazo en lugar de eventos separados. El resultado se acumula a lo largo del tiempo, produciendo una inmensa base de conocimiento que puede ser visualizada usando nuevas técnicas que permiten un acceso a la información más exhaustivo.
Palabras clave: Flujos de noticias, Extracción de eventos cross-lingual, historias

Abstract: The European project NewsReader develops advanced technology to process daily news streams in 4 languages, extracting *what* happened, *when* and *where* it happened and *who* was involved. NewsReader reads massive amounts of news coming from thousands of sources. It compares the results across sources to complement information and determine where the different sources disagree. Furthermore, it merges current news with previous news, creating a long-term history rather than separate events. The result is cumulated over time, producing an extremely large knowledge base that is visualized using new techniques to provide more comprehensive access.
Keywords: news streams, cross-lingual event extraction, history lines

1 Introduction

Professionals in any sector need to have access to accurate and complete knowledge to take well-informed decisions. Decision-makers are involved in a constant race to stay informed and to respond adequately to any changes, developments and news. However, the volume of news and documents provided by major information brokers has reached a level where state-of-the-art tools no longer provide a solution to these challenges.

The NewsReader project[1] (Vossen et al., 2014) analyzes news articles in 4 languages (English, Dutch, Italian and Spanish) to extract *what* happened, *where* and *when* it happened, and *who* was involved. NewsReader will reconstruct and visualize coherent storylines in which new events are related to past events. The system will not forget any detail, will keep track of all the facts and will even know when and how different sources told stories differently. The project will be tested on economic-financial news.

2 Objectives

One of the main goals of NewsReader is to extract events from multilingual news and to organize these events into coherent narrative storylines. The project will extract cross-lingual event information, participants taking part in the event, and additional time and location constraints. NewsReader will also detect the factuality of the events and their provenance. In addition, NewsReader will merge the news of today with previously stored information, creating a long-term his-

[1]NewsReader is funded by the European Union as project ICT- 316404. It is a collaboration of 3 European research groups and 3 companies: LexisNexis, ScraperWiki and Synerscope. The project started on January 2013 and will last 3 years. For more information see: `http://www.newsreader-project.eu/`

ISSN 1135-5948

tory rather than storing separate events. The final output will be stored in the KnowledgeStore that supports formal reasoning and inferencing.

The project foresees an estimating flow of 2 million news items per day and the complex linguistic analysis of those documents needs to be done in a reasonable time frame. The project faces thus an important challenge also regarding the scalability of the linguistic text processing.

In the same way, the amount of data produced in NewsReader is extremely large and complex. The content of the KnowledgeStore has to be offered to professional decision-makers in an effective way. NewsReader will develop innovative visualization techniques for events, their internal structure and their relations to other events that will graphically and adequately display the content of the KnowledgeStore. The visualizations of these storylines are expected to be more efficient and provide a more natural summarization of the changing world with more explanatory power.

3 Work Plan

The research activities conducted within the NewsReader project strongly rely on the cross-lingual detection of events, which are considered as the core information unit underlying news. The research focuses on four challenging aspects: event detection (addressed in WP04), event processing (addressed in WP05), storage and reasoning over events (addressed in WP06), and scaling to large textual streams (addressed in WP2). IXA group[2] is leading both WP2 and WP4.

The overall approach for processing data follows a sequence of steps, covered by the different work packages. The industrial partners define and collect relevant data sources, which are used as input by the system. The textual sources defined in WP01 (User Requirements) come in various formats.

The pieces of news are first processed through a language processing pipeline to detect event mentions, their participants and their location and time. This processing is document-based and the results are stored in the Natural Language Processing format (NAF, (Fokkens et al., 2014)). NAF is a sequel of the KYOTO Annotation Framework (KAF, (Bosma et al., 2009)) and is compliant with the Linguistic Annotation Format, LAF (Ide, Romary, and Éric Villemonte de La Clergerie, 2003).

Next, the event mentions within and across documents are compared to decide whether they refer to the same event. To represent these instances we use the Simple Event Model, SEM (Van Hage et al., 2011), which is an RDF-compliant model for representing events. Coreference can be applied to entities and to events and it can involve mentions within the same document and across documents. If different event mentions refer to the same event, duplication, complementing information and inconsistencies have to be detected. These comprise participants, place and time relations. If they make reference to different events, it is also necessary to determine the relation between them such as temporal or causal relations.

The final output, represented as NAF and SEM, is stored in the KnowledgeStore. The KnowledgeStore has different components for different types of data. It allows to store in its three interconnected layers all the typologies of content that have to be processed and produced when dealing with unstructured content and structured knowledge. The KnowledgeStore acts as a "history-recorder" which keeps track of the changes in the world as told in the media. It represents the information in RDF and supports reasoning over the data.

The next sections will describe the event detection and scalability tasks in more detail.

4 Event Detection

NewsReader uses an open and modular architecture for event detection. The system uses NAF as the layered annotation format, and separate modules have been developed to add new interpretation layers using the output of previous layers. Text-processing requires basic and generic NLP steps such as tokenization, lemmatization, part-of-speech tagging, parsing, word sense disambiguation, named-entity and semantic role recognition, etc. for all the languages within the project. Named entities are linked as much as possible to external sources such as DBpedia entity identifiers. We are also developing new techniques and resources to achieve interoperable semantic interpretation for English, Dutch, Spanish and Italian thanks to

[2] http://ixa.si.ehu.es/Ixa

the Predicate Matrix (López de Lacalle, Laparra, and Rigau, 2014).

Semantic interpretation involves the detection of event mentions and those named entities that play a role in these events, including time and location relations. This implies covering all expressions and meanings that can refer to events, their participating named entities, place and time relations. It also means to resolve coreference relations for these named entities and relations between different event mentions. As a result of this process, the text is enriched with semantic concepts and identifiers that can be used to access lexical resources and ontologies. For each unique event, we will also derive its factuality score based on the textual properties and its provenance.

NewsReader provides an abstraction layer for large-scale distributed computations, separating the *what* from the *how* of computation and isolating NLP developers from the details of concurrent programming. Section 4.1 explains the modules developed to perform event detection. Section 4.2 presents the implemented scaling infrastructure for advanced NLP processing.

4.1 NLP pipeline

We have defined a linguistic processing pipeline to automatically detect and model events. The NLP pipeline consists of basic and generic NLP processing steps, such as tokenization, lemmatization, part-of-speech tagging, word sense disambiguation and named-entity recognition. It also includes more sophisticated modules that deal with nominal coreference, nominal and verbal semantic role recognition, time recognition and interpretation, opinion detection, factuality detection, event classification and provenance identification.

Each task is executed by one independent module, which allows custom pipelines for text processing. We have developed a set of NLP tools which we refer to as the IXA pipeline (Agerri, Bermudez, and Rigau, 2014) for English and Spanish. The IXA pipeline currently provides the following linguistic annotations: Sentence segmentation, Tokenization, Part of Speech (POS) tagging, Lemmatization, Named Entity Recognition and Classification (NER), Syntactic Parsing and Coreference Resolution. This basic pipeline has been enhanced by adding new modules for word sense disambiguation, named-entity disambiguation, semantic role labeling, recognition of temporal expressions, factuality recognition, opinion mining and event coreference resolution.

The interoperability among the modules is achieved by using NAF as a common format for representing linguistic information. All the modules of the pipeline are adapted to read and write NAF, adding new layers to the NAF representation. The output can be streamed to the next module or it can be stored in the KnowledgeStore.

4.2 Scalability

The processing of news and documents provided by LexisNexis (one of the industrial partners of NewsReader and a large international news broker), has become a major challenge in the project. We have thus defined a new distributed architecture and technology for scaling up text analysis to keep pace with the rate of the current growth of news streams and collections. Scalable NLP processing requires parallel processing of textual data. The parallelization can be effectively performed at several levels, from deploying copies of the same LP among servers to the reimplementation of the core algorithms of each module using multi-threading, parallel computing. This last type of fine-grained parallelization is clearly out of the scope of the present work, as it is unreasonable to expect it to reimplement all the modules needed to perform such a complex task as mining events. We rather aim to process huge amounts of textual data by defining and implementing an architecture for NLP which allows the parallel processing of documents.

With this aim, we have created one Virtual Machine (VM) per language and pipeline so that a full processing chain in one language can be run on a single VM. This approach (Artola, Beloki, and Soroa, 2014) allows us to scale horizontally (or scale out) as a solution to the problem of dealing with massive quantities of data. We thus scale out our solution for NLP by deploying all the NLP modules into VMs and making as many copies of the VMs as necessary to process an initial batch of documents on time.

The modules are managed using the Storm framework for streaming computing[3]. Storm is an open source, general-

[3] http://storm.incubator.apache.org/

purpose, distributed, scalable and partially fault-tolerant platform for developing and running distributed programs that process continuous streams of data. Storm allows to set scalable clusters with high availability using commodity hardware and minimizes latency by supporting local memory reads and avoiding disk I/O bottlenecks.

Inside the VMs, each LP module is wrapped as a node inside the Storm topology. When a new document arrives, the processing node calls an external command sending the document to the standard input stream. The output of the LP module is received from the standard output stream and passed to the next node in the topology. Each module thus receives a NAF document with the (partially annotated) document and adds new annotations onto it. The tuples in our Storm topology comprise two elements, a document identifier and the document itself, encoded as a string with the XML serialization of the NAF document.

This setting has allowed the project to process more than 100.000 documents from the financial and economic domains using 8 copies of the VMs distributed among the project partners. As a result from the linguistic processing, more than 3 million events have been extracted.

5 Concluding Remarks

In this paper, we outlined the main objectives and methodology of the NewsReader project. We designed and implemented a complex platform for processing large volumes of news in different languages and storing the result in a KnowledgeStore that supports the dynamic growth and reasoning over data. The project shows that it is possible to develop reasoning technologies on top of the data that is generated from raw text.

Acknowledgments

This work has been supported by the EC within the 7th framework programme under grant agreement nr. FP7-IST-316040.

References

Agerri, Rodrigo, Josu Bermudez, and German Rigau. 2014. IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Ninth conference on International Language Resources and Evaluation (LREC-2014), 26-30 May, Reykjavik, Iceland.*

Artola, Xabier, Zuhaitz Beloki, and Aitor Soroa. 2014. A stream computing approach towards scalable NLP. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014).*

Bosma, Wauter, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation.*

Fokkens, Antske, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, German Rigau, Willem Robert van Hage, and Piek Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *To appear in Proceedings of 10th Joint ACL/ISO Workshop on Interoperable Semantic Annotation (ISA-10).*

Ide, Nancy, Laurent Romary, and Éric Villemonte de La Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems (SEALTS).* Association for Computational Linguistics.

López de Lacalle, Maddalen, Egoitz Laparra, and German Rigau. 2014. Predicate matrix: extending semlink through wordnet mappings. In *Ninth conference on International Language Resources and Evaluation (LREC-2014), 26-30 May, Reykjavik, Iceland.*

Van Hage, W.R., V. Malaisé, G.K.D. De Vries, G. Schreiber, and M.W. van Someren. 2011. Abstracting and reasoning over ship trajectories and web data with the simple event model (SEM). *Multimedia Tools and Applications*, pages 1–23.

Vossen, Piek, German Rigau, Luciano Serafini, Pim Stouten, Francis Irving, and Willem Van Hage. 2014. Newsreader: recording history from daily news streams. In *Ninth conference on International Language Resources and Evaluation (LREC-2014), 26-30 May, Reykjavik, Iceland.*