CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 0000; **00**:1–26
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe

NLPHub: An e-Infrastructure-based Text Mining Hub

Gianpaolo Coro*, Giancarlo Panichi,
Pasquale Pagano, Erico Perrone

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – CNR, Pisa, Italy

SUMMARY

Text mining involves a set of processes that analyse text to extract high-quality information. Among its large number of applications, there are experiments that tackle big data challenges using complex system architectures. However, text mining approaches are neither easy to discover and use nor easily combinable by end-users. Further, they should be contextualised within new approaches to Science (e.g. Open Science) that ensure longevity and re-use of methods and results.

This paper presents NLPHub, a distributed system that orchestrates and combines several state-of-the-art text mining services that recognise spatiotemporal events, keywords, and a large set of named entities. NLPHub adopts an Open Science approach, which fosters the reproducibility, repeatability, and re-usability of methods and results, by using an e-Infrastructure supporting data-intensive Science. NLPHub adds Open Science-compliance to the connected services through the use of representational standards for services and computations. It also manages heterogeneous service access policies and enables collaboration and sharing facilities. This paper reports a performance assessment based on an annotated corpus of named entities, which demonstrates that NLPHub can improve the performance of the single integrated processes by cleverly combining their output.

Copyright © 0000 John Wiley & Sons, Ltd.

Received . . .

KEY WORDS: Cloud Computing, Natural Language Processing, Software-as-a-Service, Text Processing, Text Mining, e-Infrastructures, Web Processing Service, Open Science

1. INTRODUCTION

Extracting information from a text allows knowledge to be derived from it automatically. Useful patterns and fragments of text can be detected and then re-used in a number of applications, for example to identify documents potentially relevant to a certain topic [1, 2], to give structure to unstructured information [3, 4], to produce summarised knowledge from a large quantity of documents [5, 6], to extract the concepts and topics treated by a text and to find relationships between them [7, 8]. Increasing text processing performance and usage is considered one of the future drivers of scientific progress, with immediate benefits to health and industry [9]. In fact, text processing is used in life science to summarise important results from very large collections of published documents and to apply these results in clinical trials and drug monitoring [10, 11, 12, 13, 14]. *Text mining* is the term used to indicate text processing that extracts high-quality information from a text. Text mining can be used to discover links between different studies, e.g. between different diseases [15, 16, 17]. Applications of text mining include: (i) improving text understanding [8], (ii) extracting the opinion of a group of people on a certain topic [18], (iii)

*Correspondence to: Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI) – CNR
Via G. Moruzzi, 1 – 56124, Pisa – Italy
E-mail: coro@isti.cnr.it

Copyright © 0000 John Wiley & Sons, Ltd.

categorizing documents in large catalogues [19], (iv) supporting virtual assistants and chatter-bots [20, 21, 22], and (v) automatically populating ontologies [23, 24, 25].

Today, text mining faces newly developed approaches to Science that address the challenges introduced by big data [26], i.e. managing large volumes, high production rate, heterogeneous complexity, and unreliable content. Also, these approaches try to ensure the longevity of data and methods through their re-use in complex models and process chains. Open Science (OS) is one of these paradigms, which fosters technological and methodological approaches based on the three "R"s of the scientific method: Reproducibility, Repeatability, and Re-usability [27, 28, 29]. OS suggests using collaborative systems based on Web services that support data-intensive science and the open publication of processes and results. While big data challenges are long known in text mining [30, 31], the benefits of publishing and re-using methods, and reproducing the results found by other scientists have been recognised only recently [32]. Most approaches are based on domain- or topic-specific methods that may address the repeatability and reproducibility of experiments but are difficult to re-use across domains, and thus are not OS-compliant [33, 34]. Few examples of multi-services integrations exist that build upon the complementarity of different text processing methods to offer multi-domain solutions [33, 35], but this is not sufficient to be OS-compliant because these services are not compliant with standards. OS-oriented initiatives have recently started to facilitate the use of text mining technologies for document preservation and cataloguing [36, 37]. These initiatives address discoverability and interoperability of tools and platforms through the use of standards, and are usually based on the same underlying technology and concepts presented in this paper or share resources with it.

E-Infrastructures (eIs) are distributed Computer Science systems, designed to support scientific processes, which can introduce text mining to Open Science. An eI is a network of hardware and software resources that allow remote users and services to collaborate and exchange information while supporting data-intensive science [38, 39]. An eI provides tools to integrate processes from several domains, to possibly transform them into services, and to connect data from heterogeneous sources. All eI resources are selectively made available to groups of scientists working together while managing data and services access policies. Key services of an eI are (i) distributed storage and parallel/distributed processing systems, (ii) secure multi-policy data access and harmonisation services, (iii) accounting and security services, (iv) data/models catalogues and integration services, and (v) data sharing and social networking facilities. Further, eIs foster (i) the representation of data and processes through standards, (ii) the interoperability between data-access, processing, and sharing services, and (iii) the tracking of experimental input, output, and parameters (*provenance*).

This paper presents an eI-based text mining system (NLPHub) that uses a cloud computing platform to orchestrate, interconnect, and combine the outputs of different text mining services and methods hosted by different providers and eIs. These processes recognise fragments of a text (*annotations*) associated with named abstract or physical objects (named entities), spatiotemporal events, and keywords. The integrated state-of-the-art text annotation processes come from different providers that have different access policies. The e-I is used to manage these policies while serving different groups of users. The cloud computing platform speeds up the processing for some methods and enables standard interfaces for all connected processes based on the Web Processing Service standard (WPS [40]) defined by the Open Geospatial Consortium.

An orchestrator algorithm harmonises and combines the outputs of all processes by producing one overall text annotation. Through the usage of standards for service description, the overall system makes the connected services and methods compliant with OS directives by enabling features like the sharing of results, the automatic tracking of computational provenance, reproducibility, and process re-use across multiple domains. This OS compliance allowed building the orchestrator algorithm efficiently by reducing the effort for its implementation. The name "NLPHub" refers to the fact that this platform is conceived to go beyond text annotations and will be extended to other text mining methods (e.g. sentiment analysis and opinion mining), and natural language processing tasks (e.g. text-to-speech and speech processing). The feasibility of this extension is guaranteed by the generality of the approach described in this paper, which strongly depends on the OS compliance of all connected processes.

Copyright © 0000 John Wiley & Sons, Ltd.

Overall, this paper answers the following research question: *Is it possible to build an online service for text mining that is (i) free-to-use, (ii) open-source, (iii) economically sustainable, (iv) multi-method, multi-domain, and multi-platform (i.e. integrates algorithm and services from multiple eIs while managing access policies), and (v) Open Science compliant?* This research question arises from practical issue of eIs that integrate different services from multiple eIs (Section 2.1).

This paper is organised as follows: Section 2 explains the general idea behind NLPHub, its architecture, the used e-Infrastructure, the connected services and methods, and the orchestrator process along with its interface. Section 3 measures the performance of the single processes and the entire NLPHub on a named entities recognition task using an annotated reference corpus. Section 4 discusses the results and draws conclusions.

2. METHOD

This section describes the concepts, the services, and the platform used in the NLPHub, i.e. (i) the Open Science-oriented e-Infrastructure and its cloud computing platform (Section 2.1), (ii) the concept of "named entity" (Section 2.2), (iii) the connected services and methods (Section 2.3), and (iv) the orchestration and output-merging process along with its Web interface (Section 2.4). An overview of the system architecture is given in Figure 1. Currently, NLPHub supports five languages: English, Italian, German, French, and Spanish, because these are the ones that have been requested by the European projects this software is involved in (i.e. [41, 42, 43]). Nevertheless, the system is extensible to cover other languages (e.g. Dutch, Portuguese, Swedish, Finnish etc.) as a result of its flexibility to manage different providers and the inclusion of methods that can be applied to many languages.

2.1. E-Infrastructure and Cloud Computing Platform

The open-source D4Science e-I was used as the underlying e-Infrastructure for the NLPHub [44, 45]. D4Science supports applications in many domains through the integration of a distributed storage system, a cloud computing platform, online collaborative tools, and catalogues of metadata and geospatial data. D4Science has low maintenance costs and a long-term sustainability plan based on a large number of European projects using it that cover several disciplines [46]. D4Science supports the creation and management of Virtual Research Environments (VREs) [47, 48]. A VRE is a Web-based environment offering applications that support collaboration between users working on the same topic while managing data and services access policies. In the D4Science VREs, social networking facilities allow communicating with the VREs members to share data, results, and information. Every user is granted access to a private online file system (the *Workspace*), based on a high-availability distributed storage system, that enables data and folders sharing functionalities. Users can subscribe to free-to-use VREs or request subscription to a private-access VRE moderator. VREs are the main functionality D4Science uses to manage heterogeneous access policies to services and data. Basically, public-access VREs provide interfaces and services that are free-to-use, whereas private-access VREs usually include on-payment or non-open services. The D4Science security and accounting facilities monitor the usage of all resources (storage, computational services, etc.) and prevent policy violations.

D4Science includes a cloud computing platform named DataMiner [49, 50], which currently makes ~400 processes available as-a-service and describes these processes under the WPS standard. A number of clients are embedded in third-party software [51, 52, 53, 54] that can interact with the DataMiner hosted processes through WPS (Figure 1). DataMiner allows the hosted processes to be parallelised for execution both on multiple virtual cores and on multiple machines organised as a cluster. In the free-to-use VREs, the DataMiner cluster is made up of 15 machines with Ubuntu 16.04.4 LTS x86 64 operating system, 16 virtual cores, 32 GB of RAM and 100 GB of disk space. The DataMiner machines are hosted by the National Research Council of Italy and the Italian Academic and Research Network (GARR). A load balancer distributes computational requests uniformly to the machines of the computational cluster. Each machine hosts a processing request queue that allows a maximum of 4 concurrent executions running on one machine. With this combination of parallel and distributed processing, DataMiner allows processing big data while enabling provenance tracking and results sharing [50]. At the end of a computation, the meta-information about the input and output data, and the parameters used (i.e. the computational provenance) are automatically saved on the D4Science Workspace and are described using the Prov-O XML ontological standard [55]. A Web interface is available for each process, which is automatically generated based on the WPS interpretation. Through this interface, users can select the data to process from the Workspace and conduct experiments based on shared folders that allow automatic sharing of results and provenance with other users. DataMiner also offers tools to integrate algorithms written in a multitude of programming languages [56]. In this paper, the term "algorithm" indicates processes written for the DataMiner system, whereas "method" indicates processes and workflows that were developed independently. Due to its integration flexibility, DataMiner hosts methods and algorithms for a wide range of domains, ranging from large database searches [57], to virtual reality [58] and computational biology [59].

D4Science with DataMiner helps to make our methodology compliant with Open Science through (i) the use of standards to represent the integrated processes, (ii) provenance tracking, (iii) results sharing, (iv) support of data-intensive science, (v) re-use of processes from one VRE to another one (i.e. from one domain to another). In particular, repeatability is managed by the possibility to share provenance between users and to make other users execute the same experiment exactly. Reproducibility is guaranteed by the possibility to slightly change the parameters of a shared experiment while using the same processes as before. Re-usability is a consequence of the provisioning of a process to multiple VREs while describing it via a recognised standard. Enabling OS compliance is an immediate added value of the NLPHub that attracts service providers. On the other hand, when connecting services external to D4Science, their availability is possibly subjected to a service level agreement in order to guarantee the high availability of the overall system.

Overall, the NLPHub relies on the services provided by D4Science to implement an Open Science approach for NLP tasks. In particular, as explained in the next sections, on the one hand, the NLPHub provides one access point to several NLP algorithms. On the other hand, it introduces a new paradigm for provisioning integrated services in D4Science: first, a number of algorithms addressing the same task are individually integrated to use the WPS standard and to produce uniform outputs; afterwards, one service endpoint is offered on top of them that seamlessly invokes the algorithms and merges their outputs. This approach is general enough to be used for several types of NLP tasks and is strongly facilitated by the underlying usage of the OS platform.

2.2. *Named Entities*

A "named entity" is an abstract or physical object to which a proper name can be associated. Named-entity recognisers (NERs) are Information Extraction processes that identify instances of entities in an unstructured text [60], e.g. Rome is an instance of the Location entity. Thus, entities are classes to which a NER assigns portions of an input text. The general term "annotation" is used in this paper to include other objects extracted by the connected processes that cannot be properly defined as entities, i.e. Events, Keywords, Tokens, and Sentences. Overall, the general goal of the NLPHub is to identify annotations in the text based on the classes listed in Table I. Some annotations in the table require an explanation because their meaning is not intuitive:

- *Geopolitical entity*: A geographical area associated with a political structure;
- *Misc*: Miscellaneous concepts that cannot be associated with any of the other classes, e.g. "Bachelor of Science";
- *Ordinal*: A word referring to a position in an ordered list, e.g. 1st, 2nd, etc.;
- *Token*: A sequence of characters in the text that represents a useful semantic unit;
- *Sentence*: A sequence of tokens that identifies a complete sentence;
- *Event*: Nouns, verbs, or phrases referring to a phenomenon occurring at a certain time and/or space;
- *Keyword*: A word or a phrase that is of great importance to understand the text content.

NERs use ontological classes to define named entities, and in the NLPHub these classes were made compliant with those used by the Stanford CoreNLP software, which is the largest set. One exception is the Geopolitical entity, which is recognised only by one NER (ItaliaNLP) because the other NERs displace its elements among Locations and Organizations.

2.3. Integrated Text Processing Methods

This section describes the text processing services and methods integrated with the NLPHub. Direct links to all mentioned services are provided in supplementary material. A common JSON format is used to report the recognised annotations of every integrated method. This format reports the detected annotations and their initial and final positions in the input text:

```
{"text": "input text",
  "$NER_1$": {
   "$annotations$":{
    "$annotation_1$":[
     {"$indexes$": [$i_1$,$i_2$]},
     {"$indexes$": [$i_3$,$i_4$]},
      ...,
     {"$indexes$": [$i_g$,$i_{g+1}$]}],
    ...,
    "$annotation_k$":[
     {"$indexes$": [$i_1$,$i_2$]},
     {"$indexes$": [$i_3$,$i_4$]},
      ...,
     {"$indexes$": [$i_t$,$i_{t+1}$]}]},
  ...,
  "$NER_m$": {
   "$annotations$":{
    "$annotation_1$":[
     {"$indexes$": [$i_1$,$i_2$]},
     {"$indexes$": [$i_3$,$i_4$]},
      ...,
     {"$indexes$": [$i_g$,$i_{g+1}$]}],
    ...,
    "$annotation_d$":[
     {"$indexes$": [$i_1$,$i_2$]},
     {"$indexes$": [$i_3$,$i_4$]},
      ...,
     {"$indexes$": [$i_f$,$i_{f+1}$]}]}}
}
```

All services and methods were integrated with DataMiner by writing a wrapping algorithm that transformed their original outputs into this format. This approach made it easier to build another algorithm on top of all the others, which orchestrated concurrent calls and finally built an overall annotated document for the input text (Section 2.4). The wrapping algorithms were integrated through the D4Science integration tool [56], which automatically transforms a process (e.g. an algorithm or even a service client) into a Web service invocable via the WPS standard, after the specification of the process input and output. Thus, integrating a new method requires developing a new wrapping algorithm and then using the D4Science integration tool within a VRE that is compliant with its access and usage policies. When integrating a method, the developer accepts the terms of use of the VRE that may request to make the method available also to other VREs. This option can be useful for developers who want to increment their users and application domains and overall fosters the OS concept of re-usability. Further, in these VREs the method developer can find data and corpora shared by the VRE communities, which allow refining the method itself.

For example, a NER for archaeological documents [61] was published in a D4Science VRE for archaeological studies of the Parthenos project [41]. This algorithm benefited from the feedback and the documents shared by the VRE users to improve its performance and to support the VRE community better. Further, since the VRE terms of use required accepting a re-usability clause for the integrated method, the NER was also proposed to other VREs focussing on historical document analysis and cultural heritage.

2.3.1. ***CoreNLP.*** The Stanford CoreNLP software [62] is an open-source toolkit to process texts using a large range of analysis tools. CoreNLP has been used in production applications [63] and supports a relatively large number of languages with respect to other text processing toolkits [64]. CoreNLP includes the following text mining methods: Part-of-speech tagging, named entity recognition, morphological parsing, and sentiment analysis. Currently, the supported languages are English, Arabic, Chinese, French, German, and Spanish. CoreNLP can run as-a-service where one service can manage multiple languages. Service clients can choose the language and other processing options through a JSON document sent via HTTP-Post. Thanks to its installation and operational flexibility, CoreNLP is suitable to operate within an e-Infrastructure. Further, since CoreNLP is open-source and easy to extend, plug-ins for other languages than the legacy ones can be found. Among these, Tint (The Italian NLP Tool) is an extension of CoreNLP for Italian and is distributed as a standalone Web service [65]. Different named entities are supported depending on the language, although Person, Location, and Sentence are always included.

The NLPHub integrates the CoreNLP as-a-service with English, German, French, and Spanish packages enabled, and the Tint service for Italian. Service instances were installed together on two distinct replicated D4Science virtual machines with 10 GB of RAM and 6 cores (Figure 1). Requests loads are equally balanced between the two services through an HAProxy instance [66]. DataMiner hosts one wrapping algorithm for each language, and each algorithm manages its service requests. Each language-specific DataMiner algorithm (i) receives an input text file and a list of entities to recognise (among those supported by the language), (ii) pre-processes the text by deleting useless characters (e.g. double and single quotes, brackets, non-UTF-8 characters, etc.), (iii) encodes the text using the UTF-8 charset, (iv) sends the text via HTTP-Post to the corresponding CoreNLP service and waits for the response, and finally (v) returns the annotation as an NLPHub-compliant JSON document.

2.3.2. ***GATE Cloud.*** GATE Cloud is a cloud service that offers on-payment text analysis methods as-a-service [67, 68]. It has been used in industrial and research applications, especially to process big data [69, 33]. GATE Cloud hosts a network of virtual services that provide text analysis methods. Also, it offers tools to add new methods that can use machine learning implementations made available through a Java-based integration framework. A legacy Information Extraction system (ANNIE) is available as-a-service to be used as a baseline text analysis tool. ANNIE can recognise entities like Person, Location, Organization, Date, Money, Percentage, and has an extension for processing *tweets* (of the Twitter social network) that recognises URL, Emoticon, and Hashtag classes. Another extension named ANNIE Measurements focusses on numeric expressions and measurements. Currently, GATE Cloud allows up to 1,200 free service calls per day. However, an agreement with the SoBigData European project allows D4Science to freely use several named entity recognition services in exchange for enabling OS-oriented features [70]. Among the accessible services, the ANNIE implementations for English, German, and French were integrated with the NLPHub (Table I). ANNIE Measurements is available for English only and was integrated too. Integration was operated within a controlled Virtual Research Environment that accounts for users' request loads and ensures fair usage of the free services. DataMiner wrapping algorithms were developed for each integrated GATE Cloud method and language (Figure 1). These algorithms manage users' requests towards GATE Cloud following the same workflow of the CoreNLP integration.

2.3.3. ***OpenNLP.*** The Apache OpenNLP library [71] is an open-source text processing toolkit that includes methods for language detection, tokenisation, part-of-speech tagging, morphological parsing, and named entity recognition. Most of these methods are based on machine learning models. An OpenNLP-based English NER is available as-a-service on GATE Cloud [72] and is included among the free-to-use services granted to D4Science. This service is able to recognise Person, Location, Organization, Date, Money, Percentage, and Time entities, and also offers tokenisation and sentence boundaries annotations (Table I). DataMiner hosts one wrapping algorithm that manages this method (Figure 1) using the same workflow schema of the other GATE Cloud methods.

2.3.4. ***ItaliaNLP.*** ItaliaNLP is a free-to-use service hosting a linguistic annotation pipeline for Italian that combines rule-based and machine learning algorithms [73, 74]. This service publishes endpoints to perform part-of-speech tagging, tokenisation, morphological parsing, lemmatisation, named entity recognition, clustering, words similarity assessment, and sentiment analysis. ItaliaNLP was developed by the Istituto di Linguistica Computazionale of the National Research Council of Italy (ILC-CNR) principally to support linguistic, cultural heritage, and e-learning applications [75, 76]. ILC-CNR hosts a balanced network of services that supports a large requests load (Figure 1). The NER service can recognize Person, Location, Geopolitical, and Organization entities. A DataMiner wrapping algorithm was developed to manage requests towards this service, after pre-processing the input text with the same workflow schema used for the CoreNLP integration. The enabling of OS-compliant functionalities within a monitored VRE was the main attracting feature for ILC-CNR to integrate ItaliaNLP with the NLPHub.

2.3.5. ***NewsReader.*** Events in a text are nouns, verbs, and phrases referring to a phenomenon occurring at a particular time and space. In order to detect events, an automatic recogniser identifies the "what", "when", "where", and "who" of a phenomenon and highlights the words containing this information. NewsReader is an advanced events recogniser released in 2014 by the NewsReader European project [77]. It can process text in English, Dutch, Italian, and Spanish. The recognizing method is a formal inferencing engine based on a large ontological knowledge base built upon corpora of annotated newspapers. This method processes a text and infers events while detecting their participants and time-space constraints. Finally, the method highlights the words and the phrases referring to these events. NewsReader is distributed as one virtual machine per language, containing all software required to run the process from the command line. Two balanced virtual machines were installed in D4Science for the English and Italian NewsReader versions, for a total of four machines (Figure 1). One wrapping algorithm for English and another one for Italian were developed for DataMiner to manage requests towards these machines. These algorithms (i) clean up the input text file and represent it under the required Newsreader Annotation Format (NAF), (ii) securely connect to a virtual machine via SSH protocol, (iii) run the Information Extraction process within a temporary folder, (iv) save the result and clean the folder, (v) represent the result as a JSON document for use in the NLPHub. Virtual machines for Spanish were not instantiated because this was not requested by the NLPHub users. Nevertheless, it would be straightforward to activate the described process for Spanish quickly.

2.3.6. ***TagMe.*** TagMe is a service for identifying short phrases (*spots*) in a text that can be linked to a pertinent Wikipedia page [78]. TagMe is used for text contextualisation and understanding applications in English, German, and Italian [79, 80, 81, 82]. The method augments a plain text by identifying "anchors", i.e. portions of the text that point to Wikipedia pages related to their meanings. In a first step, anchors are identified and disambiguated, i.e. for every identified anchor, only the page with the highest pertinence probability is retrieved. In a second step, the anchors are "pruned", i.e. for every anchor occurrence, only those that really refer to the Wikipedia page, given their context, are kept. D4Science already hosts the original TagMe instances on two balanced Virtual Machines with 32 GB RAM and 16 cores (Figure 1), whose average load was of 20,000 requests per month in 2018. On top of the TagMe RESTful APIs a DataMiner wrapping algorithm was built to (i) extract

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

anchors from the text, (ii) produce a JSON document for the NLPHub, and (iii) add OS-oriented features to the original services.

The anchors extracted by TagMe are words having a recognised meaning within their context. For the scopes of the NLPHub, these anchors were interpreted as keywords that can help contextualising and understanding the text. Thus, the NLPHub uses the TagMe-DataMiner algorithm as a Keyword class annotator for English, Italian, and German (Table I).

2.3.7. ***Keywords NER.*** Keywords is an open-source statistical method that produces tag clouds of words and nouns [83]. It has been used also in the H-Care award-winning human digital assistant [84]. The input of the method is made up of a text file and the indication of the text language. Tag clouds are extracted through a statistical analysis of the part-of-speech (POS) tags. The free TreeTagger software [85] is used as POS tagger because it covers 23 languages and thus makes Keywords widely applicable. In the following, the algorithm used for tag clouds extraction is reported:

Algorithm 1 Keywords - Tag Cloud

- Run TreeTagger to extract stemmed verbs, nouns, and tokens;
- Remove stop-words;
- Collect and process nouns and verbs separately;
- Record the occurrence frequency (OF) of each noun (and verb) across the tokens;
- Calculate the geometric mean (GM) and the log-normal standard deviation (LNSD) of the frequencies of nouns (and verbs);
- Select those nouns (and verbs) having occurrence frequency with distance lower than $1.5 \cdot LNSD$ from the geometric mean, i.e. those with $|OF - GM| < 1.5 \cdot LNSD$;
- Produce tag cloud of selected nouns and verbs with widths proportional to their occurrence frequencies.

The use of a log-normal distribution is due to the empirical hypothesis that the distribution of the occurrence frequencies of meaningful words across a document is similar to that of many natural systems [86]. For the scopes of the NLPHub, a DataMiner algorithm (Keywords NER) was produced that invokes Keywords directly on a DataMiner cluster machine (Figure 1) and internally extracts a tag cloud of nouns. These nouns are interpreted as keywords and are reported in a JSON document. Indeed, after heuristic tests performed on annotated corpora (Section 3), we observed that the sequence of nouns extracted by Keywords is often sufficient for a user to understand the topics treated by a text.

2.3.8. ***Language Identifier.*** The NLPHub users generally know the language of the text to analyse and provide this information when interacting with the system, via either the Web GUI or the service (Section 2.4.2). However, the NLPHub also provides a language identification process should language information not be specified. In particular, language recognition was developed as a DataMiner algorithm that had to satisfy the requirements of (i) being easily and quickly extendible to new languages, (ii) being fast, and (iii) having acceptable recognition performance. The algorithm is based on an empirical behaviour of TreeTagger (common to many POS taggers): When it is initialised on a certain language, but it processes a text written in another language, TreeTagger tends to detect many more nouns and unstemmed words than verbs and other lexical categories. The language identification algorithm works as follows:

Algorithm 2 Language Identifier

```
• Select a maximum of two sentences from the input text;
• Run parallel instances of TreeTagger on the extracted text, each initialised on one different language ($l$);
• For each language $l$:
    – Remove stop-words;
    – Calculate the ratio between nouns and non-nouns ($N\_NonN_l$);
    – Calculate the ratio between the number of tokens and the number of unidentified tokens ($T\_UID_l$);
    – Calculate the ratio between stemmed tokens and all tokens ($S\_T_l$);
    – Calculate language score as $LS_l = N\_NonN_l \cdot T\_UID_l \cdot S\_T_l$;
• Classify the text's language as the one having the highest score, i.e. $l_{opt} = \arg\max_l(LS_l)$.
```

This algorithm is applicable to all languages supported by TreeTagger and can run on every instance of DataMiner (Figure 1). The algorithm was embedded within the system since it showed an accuracy of 95% (i.e. correctly recognised files over the total number of files) on 100 sample text files covering the five languages currently supported by the NLPHub.

2.4. NLPHub

On top of the methods and services described so far, an alignment-merging algorithm orchestrates the computations and assembles the outputs. In the following sections, this algorithm is described along with the Web interface offered to the NLPHub users.

2.4.1. ***Alignment-Merging Algorithm.*** The orchestrator algorithm (AMERGE) is a DataMiner process that receives as input (i) a user-provided text, (ii) the indication of the text language (optionally), and (iii) a set of annotations to extract (selected among those supported for that language). No further algorithm parametrisation is supported, because the connected algorithms do not allow for parameter tuning as this would require provider-specific information that is mostly private (e.g. the machine learning models used by the providers). AMERGE uses a pool of 16 threads to concurrently invoke the text processing algorithms via WPS with appropriate input. The DataMiner internal queues and the HAProxy load balancers regulate the load on the D4Science machines (Figure 1). The process uses retry mechanisms to avoid issues due to random unavailability of the services or network delays. In the end, the algorithm collects the JSON documents coming from all text mining algorithms. For each requested annotation, intervals are extracted from the JSON documents, and an alignment-merging algorithm reassembles them and produces one overall sequence:

Algorithm 3 Alignment-Merging Algorithm - AMERGE

```
• For each annotation $E$:
    – Collect all annotations detected by the algorithms, i.e. all intervals with their start and end positions in the text;
    – Sort the intervals by their start position;
    – For each segment $s_i$:
        ∗ For each segment $s_j$:
            · If $s_j$ is properly included in $s_i$, process the next $s_j$;
            · If $s_i$ does not intersect $s_j$, break the loop;
            · If $s_i$ intersects $s_j$, create a new segment $su_i$ as the union of the two segments → substitute $su_i$ to $s_i$ and restart the loop on $s_j$;
        ∗ Save $s_i$ in the overall list of merged intervals $S$;
    – Associate $S$ to $E$;
• Return all $(E, S)$ pairs sets.
```

The output of the orchestrator is a JSON document in the NLPHub format reporting the aligned-merged annotations, plus a number of plain text files - one for each annotation - with square brackets delimiting the intervals.

Building the AMERGE algorithm outside of the NLPHub would have required additional effort to (i) harmonise the outputs of the connected algorithms, (ii) interface to many services and local processes, (iii) run the processes efficiently in parallel or distributed fashion. The WPS standardisation, the wrapping algorithms that make the produced annotations uniform, and the D4Science OS features allowed to overcome these issues and to strongly reduce the implementation time of AMERGE. Thus, the AMERGE process is strongly dependent on the paradigm implemented by the NLPHub. Further, these features are crucial also to build AMERGE processes for future extensions of the NLPHub to other NLP tasks (e.g. speech recognition, entity linking, opinion mining, etc.). Differently from other orchestrators of text mining processes (e.g. the BioCreative Meta-Server [87, 88, 89]), the aim of AMERGE is not to outperform the single connected methods. Instead, AMERGE uses the minimum amount of information returned by the integrated algorithms to support cases when there is no prior knowledge of the algorithms to use (Section 3). Further, AMERGE allows users to select just the algorithms they consider the most suited to a specific application domain, both before and after the process. This approach is a consequence of the requirement that AMERGE should work with multiple domains and community-provided methods in different Virtual Research Environments, without the need to re-implement or re-adapt the algorithm (cost-effectiveness).

2.4.2. ***NLPHub Interface and Service.*** The AMERGE algorithm is published on DataMiner as-a-service and has a WPS interface. This interface allows clients to invoke the process via HTTP-Post and HTTP-Get requests (Figure 1). In order to invoke this service, the client should specify an authorisation code in the HTTP request that identifies a user and a VRE [53]. The available annotations will depend on the VRE. An additional service (NLPHub-Info) allows retrieving the list of supported entities for a VRE given a user's authorisation code.

The NLPHub is endowed with a Web interface operating on top of the alignment-merging process (Figure 2). A language selection box allows the user to indicate the language of the text. The "Upload" button allows importing text from the local file system. The upload operation also performs language identification and automatically selects the identified language from the drop-down menu. The supported annotations for each language are reported in the bottom panel. The "Analyse" button executes the AMERGE algorithm, and the result is reported in a subsequent screen, where the detected annotations are highlighted in the right-hand panel. Annotations having no occurrence are coloured in grey. The left-hand panel highlights the entities found in the text and,

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

over the panel, the overall number of detected annotations is reported. A link allows downloading the JSON file produced by the AMERGE algorithm, which also contains the annotations produced by all integrated text processors. The green "Algorithms" bar allows viewing the algorithms that recognised the displayed annotations. The "Back" button returns to the initial panel.

A public version of the interface[†] tracks users' operations based on the IP addresses and uses the resources of a public-access Virtual Research Environment behind the scenes. Statistics on the resources' usage are periodically reported to the service providers. Private-access VREs in D4Science that include the NLPHub interface[‡], grant access only to a limited number of services (and thus annotations) depending on the VRE policies. For example, the GATE Cloud and ItaliaNLP services cannot be offered in commercial VREs or in VREs dedicated to companies.

3. RESULTS

3.1. Performance

The NLPHub performance was calculated by measuring the overlap and complementarity between the integrated services and the merged result. The I-CAB corpus [90], which is a named entities-annotated collection of Italian newspaper articles, was used to this aim. I-CAB was created by Fondazione Bruno Kessler and was used in the Evalita 2009 conference for a NER challenge [91, 92]. The corpus contains 527 documents manually annotated with the following entities: Person, Location, Organization, Geopolitical entity. The NLPHub was used to process I-CAB and to annotate the same entities plus Keywords. The choice of including Keywords was due to their possible interpretation as generic entities. This option is suited for users who do not know *a priori* which annotations they want to extract. Further, comparing keywords with manual annotations gives an indication of how much the extracted keywords are associated with important corpus entities.

The methods involved in the performance calculations (all focussing on Italian) were:

- CoreNLP-Tint; annotating Persons, Locations, and Organizations;
- ItaliaNLP; annotating Persons, Geopolitical entities, Locations, and Organizations;
- Keywords NER; annotating Keywords;
- TagMe; annotating Keywords.

For every supported entity, the AMERGE algorithm was run to obtain one overall annotation. The performance was calculated in terms of precision ($\frac{True\,Positives}{True\,Positives+False\,Positives}$), recall ($\frac{True\,Positives}{True\,Positives+False\,Negatives}$), and F-measure ($2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$). In the following, performance will be analysed per entity (referring to Table II):

3.1.1. ***Person.*** The highest performance according to F-measure is achieved by CoreNLP-Tint (85%), followed by AMERGE (84%), and ItaliaNLP (79%). In terms of precision, the performance ranking is still CoreNLP-Tint (78%), AMERGE (74%), and ItaliaNLP (74%). The recall ranking is AMERGE (96%), CoreNLP-Tint (93%), and ItaliaNLP (84%). Overall, the AMERGE performance is high and closer to the highest one and its recall is the highest. This is the result of the fact that the algorithm includes complementary intervals from the connected NERs and thus it extracts more valuable entities overall.

3.1.2. ***Geopolitical entities.*** The only NER supporting this entity is ItaliaNLP, whose performance is moderately high (77%). As a consequence, on this entity the AMERGE performance coincides with that of ItaliaNLP. Geopolitical entities often overlap with Locations, in fact there is high overlap with the Locations identified by CoreNLP-Tint (73% F-measure, 63% precision) and AMERGE (72% F-measure, 62% precision). Notably, AMERGE has the highest recall when Keywords are

[†] `https://nlp.d4science.org/hub/`

[‡] e.g. `https://services.d4science.org/group/parthenos_lab/nlp-hub`

interpreted as Geopolitical entities. This means that the merge of the two Keyword methods finds overall few false negatives with respect to true positives. In other words, AMERGE produces many valuable keywords that can be associated with other categorised entities.

3.1.3. ***Location.*** The NER with the highest performance on Locations is ItaliaNLP (59% F-measure), followed by AMERGE (31%), and CoreNLP-Tint (30%). The precision ranking is the same (52%, 18%, and 19% respectively), whereas recall inverts the ranking as AMERGE (88%), CoreNLP-Tint (84%), and ItaliaNLP (69%). The low performance of ItaliaNLP is a consequence of the higher tendency in I-CAB to classify Locations as Geopolitical entities, whereas ItaliaNLP has the opposite tendency.

3.1.4. ***Organization.*** The optimal NER for Organizations is CoreNLP-Tint (65%), followed by AMERGE (63%), and ItaliaNLP (58%). The precision ranking is CoreNLP-Tint (53%), ItaliaNLP (52%), and AMERGE (49%). The recall ranking is AMERGE (87%), CoreNLP-Tint (83%), and ItaliaNLP (66%). Here too, AMERGE produces more valuable annotations than the single connected NER methods, which increases its recall.

3.1.5. ***All entities.*** All annotated entities were merged into one generic entity (All) in order to simulate unknown but important information to be extracted from the corpus. An interesting property emerging from this case is that AMERGE has the highest F-measure and recognises All mostly as Person (62% F-measure). The same category is identified by CoreNLP-Tint (59%) and ItaliaNLP (51%) as the most overlapping to All. This result confirms that Person is the entity most annotated in I-CAB, as reported in [90]. Notably, all Keyword recognisers have moderate-good performance (40-45% F-measure), which indicates that Keywords would be a valuable source of information in the case of uncertainty about the entities to extract from the text.

3.1.6. ***Using AMERGE as a reference.*** As a further experiment, the entities extracted with AMERGE were used instead of the manually annotated I-CAB corpus. This approach aims at highlighting how much the entities extracted by the integrated methods overlap between them. In fact, performance calculation indicates that Locations are highly confused with Geopolitical entities (with 76-77% F-measure) and vice-versa (with 69% F-measure). Instead, the other entities are generally separated and this indicates that the different methods generally agree on the interpretation of the entities.

3.2. Agreement analysis

Cohen's Kappa [93] was used to further explore the agreement between the NER methods. This measure estimates the agreement between two methods on the extracted entities (fraction of true positives) with respect to co-classification by chance. This measure requires estimating the number of potentially classifiable tokens contained in the text. This number was approximated by using the tokeniser provided by the ItaliaNLP service. Because of this approximation, it is more realistic to refer to Fleiss' macro classification of Kappa ranges [94] rather than to the exact Kappa values. According to Fleiss' labels, there is a general poor agreement between the NER methods focussing on different entities, but there is high overlap (i.e. "good" agreement) between Geopolitical entities and Locations (Table III).

In some cases, AMERGE acts as an intermediary between two methods. For example, both ItaliaNLP and CoreNLP-Tint have excellent agreement with AMERGE on Organizations, but they have just "good" agreement between them. Instead, in the case of Keywords recognition, AMERGE has excellent agreement with Keywords NER, but an only marginal agreement with TagMe. This is the result of the fact that Keywords NER generally produces more entities than TagMe, which translates into a higher agreement with AMERGE.

3.3. Annotation examples

This section reports examples of successful and unsuccessful recognition of named entities by the connected methods, along with their influence on the AMERGE algorithm. One first example is the following: In the sentence "recuperati contributi per quasi 130 000 euro" (*contributions recovered for almost 130 000 euros*) from an article on economy of the I-CAB corpus, ItaliaNLP recognised "per" (*for*) as an Organisation. In this case, "per" was erroneously recognised as a specification of the type of "contribution" rather than as a preposition. Further, CoreNLP-Tint did not recognize any entity in the phrase. Thus, the error made by the ItaliaNLP NER decreased the overall precision of AMERGE on Organisations recognition. However, in the same news, ItaliaNLP correctly recognised Organisations like "Camera di commercio" (*Chamber of Commerce*) and "Organi di vigilanza" (*Supervisory bodies*) that were missed by CoreNLP-Tint, and this increased the precision of AMERGE. In the sentence "ambienti di lavoro dell'azienda per i servizi sanitari" (*work environments of the company for health services*) from the same article, ItaliaNLP recognised "azienda" (*company*) and "servizi sanitari" (*health services*) as two Organisations, whereas CoreNLP-Tint identified "azienda per i servizi" (*company for services*) as one Organisation. As a result, AMERGE reported the merged entity "azienda per i servizi sanitari" (*company for health services*) as an Organisation, which is more correct than the entities identified by the two other processes separately.

In an I-CAB article about the psychological consequences of war on Russian and Chechen children, ItaliaNLP was able to detect all Persons involved in the news, i.e. "Pirjo Honkasalo", "Putin", and "Hadizhat Gataeva", whereas CoreNLP-Tint did not detect any Person entity. The found entities enriched the output of AMERGE and increased its precision with respect to CoreNLP-Tint. Interestingly, Keywords NER detected "Pirjo Honkasalo" and "Putin" as Keywords, which thus overlap with valuable named entities, in agreement with the results reported in Section 3.1.5.

The usefulness of the combination of complementary algorithms and annotations is evident on the output of the NLPhub for a poem from Dante Alighieri's Rime (Rima LXXIX - *Voi che 'ntendendo il terzo ciel movete*), written in the *Dolce Stil Novo* ancient Italian style. On this text, the only named entity detected by the NERs was "Sire" (*Lord*) as a Person, but no other entity was detected. However, several Keywords were detected and the outputs of TagMe and Keywords NER were complementary and informative. In particular, TagMe identified the main topic of the text as related to the desire to see and have contact with a woman ("donna guardare", "veder", "vide","verace","piace"), whereas Keywords NER extracted words related to love ("Amor","paura","angoscia","sospiri"), to the woman addressed by the poet ("donna","angela","ancella"), and to her qualities ("core","occhi","anima","grandezza","miracoli adornezza"). This observation is evident with the sentence "Questi mi face una donna guardare" (*This [presence] makes me look at a woman*), from which TagMe identifies "donna guardare" as the only Keyword (but associates it to a movie reported in Wikipedia) and Keywords NER detects "face", "donna", and "guardare" as separate Keywords. Overall, the merge of these complementary results is a set of Keywords that allow to understand the main focus of the poem and also compensates the poor information extracted by the other methods.

4. DISCUSSION AND CONCLUSIONS

This paper has described the NLPHub, a distributed system connecting and combining several text processing methods and services while adding Open Science-oriented features to them. The NLPHub provides one single access endpoint to a large set of Information Extraction methods for five languages. The results show that all the connected methods have high performance on specific entities, but there is not one method outperforming the others on all entities. Therefore, there are several advantages in using the NLPHub: If a user wants to extract one particular entity and knows that a method has high performance on that entity, then using this specific method would be the best choice. Alternatively, if the user does not have knowledge about the methods' performance, using AMERGE is generally the best choice, especially due to its property to act as an intermediary

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

between different methods. Although the precision of AMERGE is comparable with that of the connected methods, this algorithm has usually higher recall than the others, which is one drawback of using AMERGE instead of another algorithm specifically developed for the text's domain. On the other hand, AMERGE generally preserves high precision also when NERs developed for heterogeneous domains are connected. For example, introducing a NER trained on archaeology [61] among the set of algorithms of Section 3 does not change the AMERGE performance sensibly, also due to the low number of entities detected by the domain-specific NER. If the user does not know which are the entities to extract, the AMERGE Keywords process provides meaningful information possibly corresponding to several named entities. Additionally, the NLPHub endows the connected methods with WPS and Web interfaces, provenance management, results sharing, and access and usage policies control through Virtual Research Environments. These advantages make the original methods and services more Open Science compliant and are therefore attractive for services providers. Thus, AMERGE is a solution entangled with the concept of Open Science, which satisfies our research question and its implicit constraints, although not outperforming the connected methods.

The NLPHub is particularly suited for linguists who just want to focus on text analysis and avoid software and hardware problems. It is also a tool for automatic agents that need to extract knowledge from large texts automatically and possibly build upon the extracted information. For example, automatic ontology population systems can use the NLPHub annotations to extract semantic triples [95, 96, 23]. Also, since the NLPHub supports Event and Keyword annotations, it can be used to automatically extract narratives out of a document [97, 98], and some experiments have already used it for this purpose [99, 100].

Future extensions of the NLPHub will include the coverage of other text mining methods (e.g. sentiment analysis, opinion mining, entity linking, and morphological parsing) and additional NLP tasks (e.g. text-to-speech, and speech processing as-a-service), which are already supported by methods and services integrated with D4Science. The NLPHub implements a paradigm where all connected algorithms are integrated as WPS processes and use the same output format, and an orchestrator algorithm concurrently merges the outputs of all connected methods. In this view, most of the D4Science features used for building the AMERGE algorithm for named entity recognition (e.g. cloud computing, WPS standard, the D4Science integration system, etc.) will be directly used for the future supported NLP tasks. This operation will require developing a new AMERGE algorithm for each NLP task (for example, to solve ambiguous overlapping annotations), but the overall OS-compliant approach will still facilitate the implementation.

The features and the paradigm of the NLPHub are a novelty with respect to other solutions for text processing. For example, initiatives specifically conceived to connect huge scholarly literature (e.g. OpenAire and OpenMinTeD [36, 37]) currently do not support multiple access and sharing policies for different communities through VREs. Instead, they usually contain registries of text and data mining applications that are not orchestrated and do not foster the use of communication standards. On the other hand, workflow management systems that support the construction of text mining process orchestrators (e.g. UIMA [101]), do not support WPS and Open Science features and do not offer a sustainable and free-to-use e-Infrastructure to deploy the workflows. Other solutions that propose an interoperability standard are very tied to one domain (e.g. BioCreative Meta-Server molecular biology [89]), which does not exclude being compliant with OS directives [102], but have smaller constraints for building the orchestrators. Our Open Science and free-to-use standardised services are a crucial difference with respect to alternative on-payment solutions (e.g. GATE Cloud). Indeed, the D4Science VREs are currently used in large initiatives on text processing, like SoBigData [70], to build upon community-specific catalogues of processes that do not natively support WPS and do not have harmonised outputs (including the algorithms of GATE Cloud). Finally, valid initiatives like the European Language Grid (ELG [103]) aim at creating multi-language environments where users can register and integrate NLP services and language resources, and build complex NLP workflows. However, the NLPHub gives higher stress than ELG on (i) using standards, (ii) tracing the computational provenance, (iii) building orchestrator algorithms quickly, for services that share the same scopes, (iv) fostering a high growth rate of the whole system through the easy connection of new algorithms, (v) enabling cooperation and domain-specific algorithms in VREs, (vi) making scientific data freely accessible within communities of practice, and overall (vii) fostering Open Science across different scientific communities.

SUPPLEMENTARY MATERIAL

The services reported in this supplementary material are maintained by the D4Science e-Infrastructure (`www.d4science.org`).

The Keyword Tag Cloud Algorithm process is available as open source software at the following link:

`https://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/LatentSemanticAnalysis/`

All services/algorithms are available for use on the D4Science platform gateway (`https://services.d4science.org`) after registration to the RPrototypingLab Virtual Research Environment (VRE):

`https://services.d4science.org/group/rprototypinglab/rprototypinglab`

A user guide on how to use the processes and services, also via WPS is available at the following link:

`https://wiki.gcube-system.org/gcube/DataMiner_Manager`

The NLPHub public interface is available at `http://nlp.d4science.org/hub/`.

The DataMiner CoreNLP algorithms are available at the following links:

`https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.ENGLISH_NER_CORENLP`
`https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.FRENCH_NER_CORENLP`
`https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.GERMAN_NER_CORENLP`
`https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.ITALIAN_NER_TINT_CORENLP`
`https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.SPANISH_NER_CORENLP`

The GATE Cloud algorithms connected to NLPHub are available at the following links:

`https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.ENGLISH_NAMED_ENTITY_RECOGNIZER`
`https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.ANNIE_PLUS_MEASUREMENTS`

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
GERMAN_NAMED_ENTITY_RECOGNIZER
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
FRENCH_NAMED_ENTITY_RECOGNIZER
```

The OpenNLP algorithm is available at the following link:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
OPEN_NLP_ENGLISH_PIPELINE
```

The ItaliaNLP algorithm is available at the following link:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
ITALIANLP_NER
```

The TagMe Keywords extraction algorithms are available for 3 supported languages at the following links:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
TAGME_ENGLISH_NER
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
TAGME_GERMAN_NER
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
TAGME_ITALIAN_NER
```

Keywords NER is available for the 5 supported languages at the following links:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
KEYWORDS_NER_ENGLISH
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
KEYWORDS_NER_FRENCH
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
KEYWORDS_NER_GERMAN
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
KEYWORDS_NER_ITALIAN
```

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
KEYWORDS_NER_SPANISH
```

Newsreader is available for the 2 supported languages at the following links:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
ENGLISH_EVENTS_RECOGNITION_NER
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
ITALIAN_EVENTS_RECOGNITION_NER
```

The AMERGE algorithm is available at the following link:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
NLPHUB_NER
```

The Language Identifier algorithm is available at following link:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
LANGUAGE_RECOGNIZER
```

The NLPHub information system is available at the following link:

```
https://services.d4science.org/group/rprototypinglab/
data-miner?OperatorId=org.gcube.dataanalysis.wps.
statisticalmanager.synchserver.mappedclasses.transducerers.
NLPHUB_INFO
```

References

1. Delen D, Crossland MD. Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications* 2008; **34**(3):1707–1720.
2. Gentzkow M, Kelly BT, Taddy M. Text as data. *Technical Report*, National Bureau of Economic Research 2017.
3. Singhal A, Simmons M, Lu Z. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS computational biology* 2016; **12**(11):e1005 017.
4. Simmons M, Singhal A, Lu Z. Text mining for precision medicine: Bringing structure to ehrs and biomedical literature to understand genes and health. *Translational Biomedical Informatics*. Springer, 2016; 139–166.
5. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. *Research Synthesis Methods* 2011; **2**(1):1–14.
6. OMara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* 2015; **4**(1):5.
7. Nasukawa T, Nagano T. Text analysis and knowledge mining system. *IBM systems journal* 2001; **40**(4):967–984.
8. Gupta V, Lehal GS, *et al.*. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence* 2009; **1**(1):60–76.
9. LIBER. "text and data mining: Its importance and the need for change in europe." 2016. Association of European Research Libraries `https://libereurope.eu/wp-content/uploads/Text%20and%20Data%20Mining%20Factsheet.pdf`.
10. Ananiadou S, Kell DB, Tsujii Ji. Text mining and its potential applications in systems biology. *Trends in biotechnology* 2006; **24**(12):571–579.
11. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome biology* 2008; **9**(2):S8.

12. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome biology* 2008; **9**(6):R96.
13. Paul MJ, Sarker A, Brownstein JS, Nikfarjam A, Scotch M, Smith KL, Gonzalez G. Social media mining for public health monitoring and surveillance. *Biocomputing 2016: Proceedings of the Pacific Symposium*, World Scientific, 2016; 468–479.
14. Blankers M, van der Gouwe D, van Laar M. 4-fluoramphetamine in the netherlands: Text-mining and sentiment analysis of internet forums. *International Journal of Drug Policy* 2019; **64**:34–39.
15. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. *Methods* 2015; **74**:97–106.
16. Mahmood AA, Wu TJ, Mazumder R, Vijay-Shanker K. Dimex: a text mining system for mutation-disease association extraction. *PloS one* 2016; **11**(4):e0152 725.
17. Sarntivijai S, Vasant D, Jupp S, Saunders G, Bento AP, Gonzalez D, Betts J, Hasan S, Koscielny G, Dunham I, *et al.*. Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. *Journal of biomedical semantics* 2016; **7**(1):8.
18. Pang B, Lee L, *et al.*. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2008; **2**(1–2):1–135.
19. Sebastiani F. Text categorization. *Encyclopedia of Database Technologies and Applications*. IGI Global, 2005; 683–687.
20. Sha G. Ai-based chatterbots and spoken english teaching: a critical analysis. *Computer Assisted Language Learning* 2009; **22**(3):269–281.
21. Kuligowska K, Lasek M. Virtual assistants support customer relations and business processes. *The 10th International Conference on Information Management, Gdańsk*, 2011.
22. Chakrabarti C, Luger GF. Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications* 2015; **42**(20):6878–6897.
23. Gillani S, Ko A. Incremental ontology population and enrichment through semantic-based text mining: an application for it audit domain. *International Journal on Semantic Web and Information Systems (IJSWIS)* 2015; **11**(3):44–66.
24. Reyes-Ortiz JA, Bravo M, Pablo H. Web services ontology population through text classification. *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, 2016; 491–495.
25. Ganino G, Lembo D, Scafoglieri F. Ontology population from raw text corpus for open-source intelligence. *International Conference on Web Engineering*, Springer, 2017; 173–186.
26. James M, Michael C, Brad B, Jacques B, Richard D, Charles R, Angela H. Big data: The next frontier for innovation, competition, and productivity. *The McKinsey Global Institute* 2011; .
27. Hey T, Tansley S, Tolle KM, *et al.*. *The fourth paradigm: data-intensive scientific discovery*, vol. 1. Microsoft research Redmond, WA, 2009.
28. EU Commission. Open science (open access) 2016. `https://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access`.
29. Assante M, Candela L, Castelli D, Cirillo R, Coro G, Frosini L, Lelii L, Mangiacrapa F, Pagano P, Panichi G, *et al.*. Enacting open science by D4Science. *Future Generation Computer Systems* 2019; .
30. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management* 2015; **35**(2):137–144.
31. Amado A, Cortez P, Rita P, Moro S. Research trends on big data in marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics* 2018; **24**(1):1–7.
32. Linthicum DS. Cloud computing changes data integration forever: What's needed right now. *IEEE Cloud Computing* 2017; **4**(3):50–53.
33. Bontcheva K, Derczynski L. Extracting information from social media with gate. *Working with Text*. Elsevier, 2016; 133–158.
34. Adedugbe O, Benkhelifa E, Campion R. A cloud-driven framework for a holistic approach to semantic annotation. *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, IEEE, 2018; 128–134.
35. Wei CH, Leaman R, Lu Z. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics* 2016; **32**(12):1907–1910.
36. OpenMinTeD. Open Mining INfrastructure for TExt and Data 2019. `https://cordis.europa.eu/project/rcn/194923/factsheet/en`.
37. OpenAire. European project supporting Open Access 2019. `https://www.openaire.eu/`.
38. Pollock N, Williams R. E-infrastructures: How do we know and understand them? strategic ethnography and the biography of artefacts. *Computer Supported Cooperative Work (CSCW)* 2010; **19**(6):521–556.
39. Andronico G, Ardizzone V, Barbera R, Becker B, Bruno R, Calanducci A, Carvalho D, Ciuffo L, Fargetta M, Giorgio E, *et al.*. e-infrastructures for e-science: a global view. *Journal of Grid Computing* 2011; **9**(2):155–184.
40. Schut P, Whiteside A. OpenGIS Web Processing Service 2007. OGC project document `http://www.opengeospatial.org/standards/wps`.
41. Parthenos. The Parthenos European Project 2019. `http://www.parthenos-project.eu/`.
42. SoBigData. The SoBigData European Project 2019. `http://sobigdata.eu/index`.
43. Ariadne. The AriadnePlus European Project 2019. `https://ariadne-infrastructure.eu/`.
44. Candela L, Castelli D, Coro G, Pagano P, Sinibaldi F. Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience* 2013; .
45. Assante M, Candela L, Castelli D, Cirillo R, Coro G, Frosini L, Lelii L, Mangiacrapa F, Marioli V, Pagano P, *et al.*. The gcube system: Delivering virtual research environments as-a-service. *Future Generation Computer Systems* 2019; **95**:445–453.
46. D4Science. The D4Science e-Infrastructure Supporting Projects 2019. `https://services.d4science.org/thematic-gateways`.

Copyright © 0000 John Wiley & Sons, Ltd.
Prepared using cpeauth.cls

47. Candela L, Castelli D, Pagano P. Virtual research environments: an overview and a research agenda. *Data Science Journal* 2013; :GRDI–013.
48. Assante M, Candela L, Castelli D, Coro G, Lelii L, Pagano P. Virtual research environments as-a-service by gcube. *PeerJ Preprints* 2016; **4**:e2511v1.
49. Coro G, Candela L, Pagano P, Italiano A, Liccardo L. Parallelizing the execution of native data mining algorithms for computational biology. *Concurrency and Computation: Practice and Experience* 2015; **27**(17):4630–4644.
50. Coro G, Panichi G, Scarponi P, Pagano P. Cloud computing in a distributed e-infrastructure using the web processing service standard. *Concurrency and Computation: Practice and Experience* 2017; **29**(18):e4219.
51. OpenCPU. Producing and reproducing results 2016. `https://www.opencpu.org`.
52. ArcMap. Arcgis for desktop 2016. `http://desktop.arcgis.com/en/arcmap/`.
53. CNR. gCube WPS thin clients 2016. `https://wiki.gcube-system.org/gcube/How_to_Interact_with_the_DataMiner_by_client`.
54. QGIS. A free and open source geographic information system 2016. `http://qgis.org/en/site/`.
55. Lebo T, Sahoo S, McGuinness D, Belhajjame K, Cheney J, Corsar D, Garijo D, Soiland-Reyes S, Zednik S, Zhao J. Prov-o: The prov ontology. *W3C Recommendation* 2013; **30**.
56. Coro G, Panichi G, Pagano P. A web application to publish r scripts as-a-service on a cloud computing platform. *Bollettino di Geofisica Teorica ed Applicata* 2016; **57**:51–53.
57. Berghe EV, Coro G, Bailly N, Fiorellato F, Aldemita C, Ellenbroek A, Pagano P. Retrieving taxa names from large biodiversity data collections using a flexible matching workflow. *Ecological Informatics* 2015; **28**:29–41.
58. Coro G, Palma M, Ellenbroek A, Panichi G, Nair T, Pagano P. Reconstructing 3d virtual environments within a collaborative e-infrastructure. *Concurrency and Computation: Practice and Experience* 2018; :e5028.
59. Coro G, Vilas LG, Magliozzi C, Ellenbroek A, Scarponi P, Pagano P. Forecasting the ongoing invasion of lagocephalus sceleratus in the mediterranean sea. *Ecological Modelling* 2018; **371**:37–49.
60. Nadeau D, Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 2007; **30**(1):3–26.
61. Daniel Williams. ArchNer it - Archaeological NER web-service for Italian language text 2018. `https://services.d4science.org/group/rprototypinglab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.ARCHNER_IT_FILE`.
62. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The stanford corenlp natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014; 55–60.
63. Song M, Chambers T. Text mining with the stanford corenlp. *Measuring scholarly impact*. Springer, 2014; 215–234.
64. Stanford University. Stanford CoreNLP - Human Languages Supported 2019. `https://stanfordnlp.github.io/CoreNLP/`.
65. Aprosio AP, Moretti G. Italy goes to stanford: a collection of corenlp modules for italian. *arXiv preprint arXiv:1609.06204* 2016; .
66. Tarreau W, *et al.*. HAProxy-the reliable, high-performance TCP/HTTP load balancer 2012. `http://haproxy.lwt.eu`.
67. GATE Cloud. GATE Cloud: Text Analytics in the Cloud 2019. `https://cloud.gate.ac.uk/`.
68. Tablan V, Roberts I, Cunningham H, Bontcheva K. GATE Cloud.net: Cloud Infrastructure for Large-Scale, Open-Source Text Processing. *UK e-Science All hands Meeting*, 2011.
69. Bontcheva K, Cunningham H, Roberts I, Roberts A, Tablan V, Aswani N, Gorrell G. Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation* 2013; **47**(4):1007–1029.
70. SoBigData European Project. Deliverable D2.7 - IP principles and business models 2016. `http://project.sobigdata.eu/material`.
71. Kottmann J, Margulies B, Ingersoll G, Drost I, Kosin J, Baldridge J, Goetz T, Morton T, Silva W, Autayeu A, *et al.*. Apache OpenNLP. *`www.opennlp.apache.org`* 2011; .
72. GATE Cloud. OpenNLP English Pipeline 2019. `https://cloud.gate.ac.uk/shopfront/displayItem/opennlp-english-pipeline`.
73. ILC-CNR. The ItaliaNLP REST Service 2019. `http://api.italianlp.it/docs/`.
74. Dell'Orletta F, Venturi G, Cimino A, Montemagni S. T2kˆ2: a system for automatically extracting and organizing knowledge from texts. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.
75. Boschetti F, Cimino A, Dell'Orletta F, Lebani G, Passaro L, Picchi P, Venturi G, Montemagni S, Lenci A. Computational analysis of historical documents: An application to italian war bulletins in world war i and ii. *Workshop on Language resources and technologies for processing and linking historical documents and archives (LRT4HDA 2014)*, ELRA, 2014; 70–75.
76. Cimino A, Dell'Orletta F, Venturi G, Montemagni S. Linguistic profiling based on general–purpose features and native language identification. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 2013; 207–215.
77. Vossen P, Agerri R, Aldabe I, Cybulska A, van Erp M, Fokkens A, Laparra E, Minard AL, Aprosio AP, Rigau G, *et al.*. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems* 2016; **110**:60–85.
78. Ferragina P, Scaiella U. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). *Proceedings of the 19th ACM international conference on Information and knowledge management*, ACM, 2010; 1625–1628.
79. Suchanek F, Weikum G. Knowledge harvesting from text and web sources. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, IEEE, 2013; 1250–1253.
80. Basile P, Caputo A, Semeraro G. Entity linking for italian tweets. *Proc. Second Italian Conf. Computational Linguistics CLiC-it 2015*, Accademia University Press, 2015; 36–40.

Copyright © 0000 John Wiley & Sons, Ltd.

81. Tran NK, Ceroni A, Kanhabua N, Niederée C. Time-travel translator: Automatically contextualizing news articles. *Proceedings of the 24th International Conference on World Wide Web*, ACM, 2015; 247–250.
82. Weiland L, Hulpus I, Ponzetto SP, Dietz L. Understanding the message of images with knowledge base traversals. *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ACM, 2016; 199–208.
83. Coro G. The Keywords Tag Cloud Algorithm 2019. `https://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/LatentSemanticAnalysis/`.
84. SpeechTEK 2010. SpeechTEK 2010 - H-Care Avatar wins People's Choice Award 2019. `http://web.archive.org/web/20160919100019/http://www.speechtek.com/europe2010/avatar/`.
85. Schmid H. Treetagger - a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart* 1995; **43**:28.
86. Scheffer M. *Critical transitions in nature and society*, vol. 16. Princeton University Press, 2009.
87. Leitner F, Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, Kuo CJ, Hsu CN, Tsai RTH, Hung HC, Lau WW, *et al.*. Introducing meta-services for biomedical information extraction. *Genome biology* 2008; **9**(S2):S6.
88. Smith L, Tanabe LK, nee Ando RJ, Kuo CJ, Chung IF, Hsu CN, Lin YS, Klinger R, Friedrich CM, Ganchev K, *et al.*. Overview of biocreative ii gene mention recognition. *Genome biology* 2008; **9**(S2):S2.
89. Leitner F, Krallinger M, Alfonso V. Biocreative meta-server and text-mining interoperability standard. *Encyclopedia of systems biology* 2013; **8401**:106–10.
90. Magnini B, Pianta E, Girardi C, Negri M, Romano L, Speranza M, Bartalesi V, Sprugnoli R. I-CAB: the Italian Content Annotation Bank. *LREC*, Citeseer, 2006; 963–968.
91. Magnini B, Cappelli A, Pianta E, Speranza M, Bartalesi V, Sprugnoli R, Romano L, Girardi C, Negri M. Annotazione di contenuti concettuali in un corpus italiano: I-CAB. *Proc. of SILFI 2006* 2006; .
92. Speranza M. The named entity recognition task at evalita 2009. *Proceedings of the Workshop Evalita*, 2009.
93. Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 1960; **20**(1):37–46.
94. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological bulletin* 1971; **76**(5):378.
95. Celjuska D, Vargas-Vera M. Ontosophie: A semi-automatic system for ontology population from text. *International Conference on Natural Language Processing (ICON)*, 2004; 60.
96. Petasis G, Karkaletsis V, Paliouras G, Krithara A, Zavitsanos E. Ontology population and enrichment: State of the art. *Knowledge-driven multimedia information extraction and ontology evolution*, Springer-Verlag, 2011; 134–166.
97. Bartalesi V, Meghini C, Metilli D. A conceptualisation of narratives and its expression in the CRM. *International Journal of Metadata, Semantics and Ontologies* 2017; **12**(1):35–46.
98. Metilli D, Bartalesi V, Meghini C. A Wikidata-based tool for building and visualising narratives. *International Journal on Digital Libraries* 2019; :1–16.
99. Metilli D, Bartalesi V, Meghini C, Aloia N. Populating Narratives Using Wikidata Events: An Initial Experiment. *Italian Research Conference on Digital Libraries*, Springer, 2019; 159–166.
100. Metilli D, Bartalesi V, Meghini C. Steps towards a system to extract. *Proceedings of the Text2Story 2019 Workshop*, Springer, 2019; na.
101. Hajič jr J, Veselovská K. Uima: Unstructured information management architecture for data mining applications and developing an annotator component for sentiment analysis 2013. `https://ufal.ms.mff.cuni.cz/~veselovska/2013/docs/UIMA_Unstructured_Information_Management_Architecture_for_Data_Mining.pdf`.
102. Coro G, Trumpy E. Predicting geographical suitability of geothermal power plants. *Journal of Cleaner Production* 2020; :121 874.
103. European Language Grid. The European Language Grid Platform 2019. `https://www.european-language-grid.eu`.

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

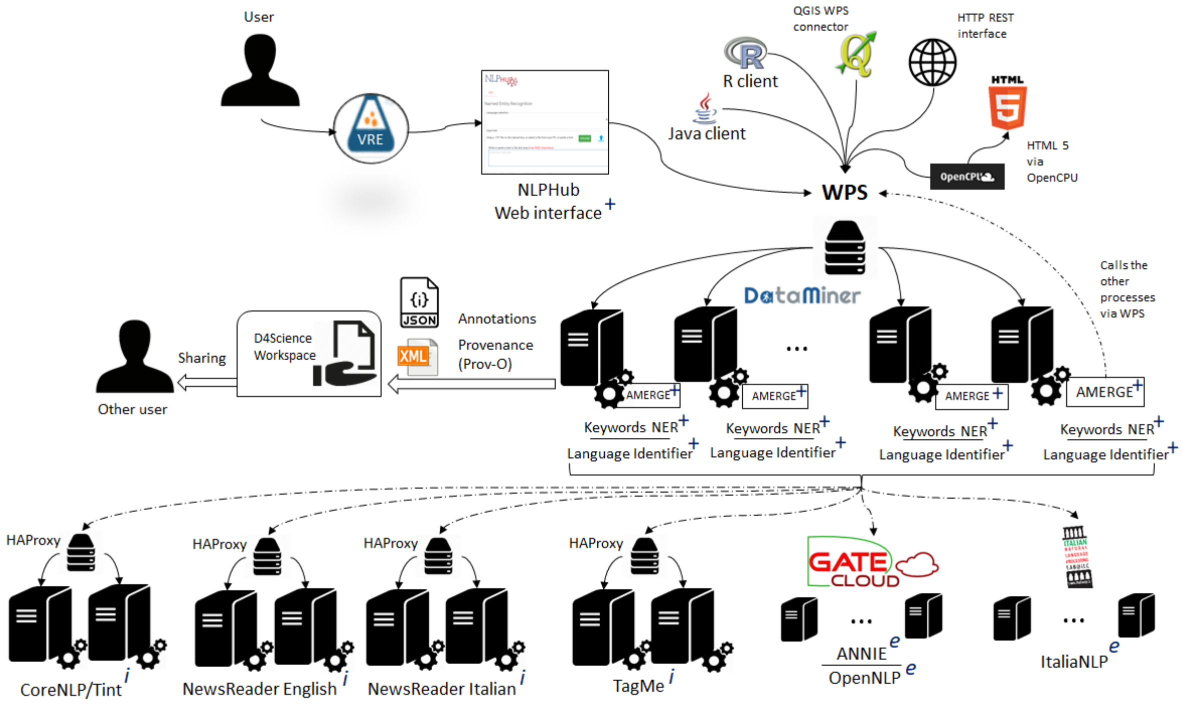
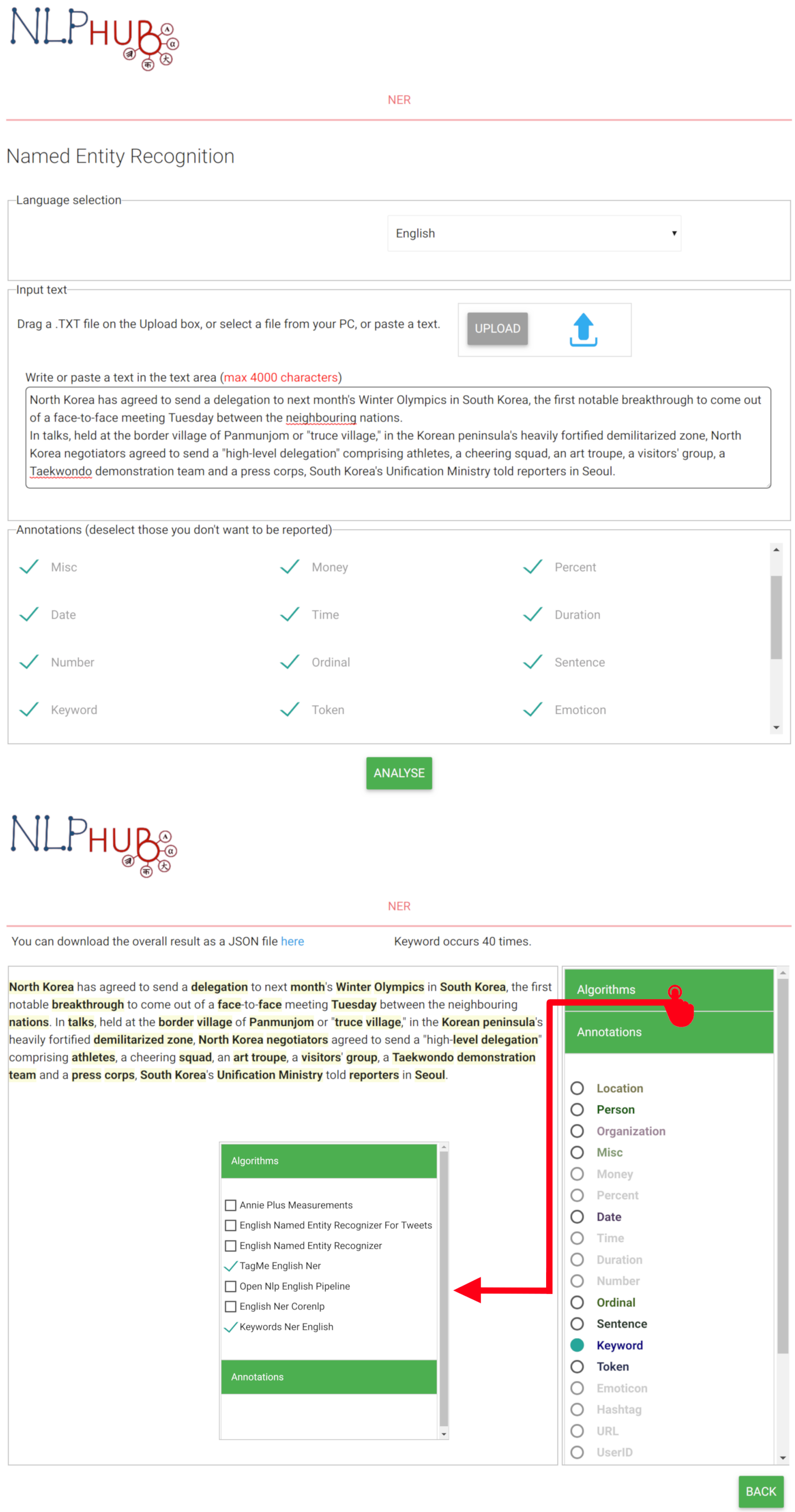


Figure 1. Architectural schema of the NLPHub. The DataMiner cloud computing system directly hosts the Keywords NER, the Language Identifier and the AMERGE algorithms. The other NLPHub algorithms running on DataMiner manage the computational requests towards text mining services. All DataMiner algorithms write their output (i.e. an annotated text in JSON format) and computational provenance on the D4Science Workspace, an online file system that enables information sharing with other users. All algorithm are endowed with a WPS description that enables a number of clients and interfaces to interact with the cloud computing platform and thus to execute the text mining algorithms. The new implemented components are indicated with the + superscript, to distinguish them from pre-existing imported methods hosted by D4Science as-a-service (i superscript), and methods hosted by services external to D4Science (e superscript).



Named Entity Recognition

Language selection

English ▾

Input text

Drag a .TXT file on the Upload box, or select a file from your PC, or paste a text.

Write or paste a text in the text area (max 4000 characters)

North Korea has agreed to send a delegation to next month's Winter Olympics in South Korea, the first notable breakthrough to come out of a face-to-face meeting Tuesday between the neighbouring nations.
In talks, held at the border village of Panmunjom or "truce village," in the Korean peninsula's heavily fortified demilitarized zone, North Korea negotiators agreed to send a "high-level delegation" comprising athletes, a cheering squad, an art troupe, a visitors' group, a Taekwondo demonstration team and a press corps, South Korea's Unification Ministry told reporters in Seoul.

Annotations (deselect those you don't want to be reported)

✓ Misc	✓ Money	✓ Percent
✓ Date	✓ Time	✓ Duration
✓ Number	✓ Ordinal	✓ Sentence
✓ Keyword	✓ Token	✓ Emoticon

You can download the overall result as a JSON file here

Keyword occurs 40 times.

North Korea has agreed to send a delegation to next month's Winter Olympics in South Korea, the first notable breakthrough to come out of a face-to-face meeting Tuesday between the neighbouring nations. In talks, held at the border village of Panmunjom or "truce village," in the Korean peninsula's heavily fortified demilitarized zone, North Korea negotiators agreed to send a "high-level delegation" comprising athletes, a cheering squad, an art troupe, a visitors' group, a Taekwondo demonstration team and a press corps, South Korea's Unification Ministry told reporters in Seoul.

Algorithms

- □ Annie Plus Measurements
- □ English Named Entity Recognizer For Tweets
- □ English Named Entity Recognizer
- ✓ TagMe English Ner
- □ Open Nlp English Pipeline
- □ English Ner Corenlp
- ✓ Keywords Ner English

Annotations

Algorithms

Annotations

- ○ Location
- ○ Person
- ○ Organization
- ○ Misc
- ○ Money
- ○ Percent
- ○ Date
- ○ Time
- ○ Duration
- ○ Number
- ○ Ordinal
- ○ Sentence
- ● Keyword
- ○ Token
- ○ Emoticon
- ○ Hashtag
- ○ URL
- ○ UserID

Figure 2. Web interface of the NLPHub. The top image reports the initial selection panel, where the user indicates the text or file to process and the annotations to detect. The lower image reports the result, with all the detected annotations highlighted with colours in the right-hand side panel. The words corresponding to the selected annotation are highlighted in the text in the left-hand panel. By pressing the "Algorithms" bar, the methods that identified the selected annotation are highlighted and can be selectively disabled together with their related annotations.

		Annotations																			
Language	**Service**	**Person**	**Location**	**Geopolitical**	**Organization**	**Date**	**Money**	**Percentage**	**Address**	**Misc**	**Keyword**	**Event**	**Number**	**Ordinal**	**Time**	**Duration**	**URL**	**Emoticon**	**Hashtag**	**Token**	**Sentence**
English	CoreNLP	✓	✓		✓	✓	✓	✓		✓			✓	✓	✓	✓					✓
	GATE Cloud - ANNIE	✓	✓		✓	✓	✓	✓	✓								✓	✓	✓	✓	✓
	GATE Cloud - ANNIE Measurements	✓	✓		✓	✓	✓	✓	✓											✓	✓
	OpenNLP	✓	✓		✓	✓	✓	✓							✓					✓	✓
	NewsReader											✓									
	TagMe										✓										
	Keywords NER										✓										
Italian	CoreNLP - Tint	✓	✓		✓																✓
	ItaliaNLP	✓	✓	✓	✓																
	NewsReader											✓									
	TagMe										✓										
	Keywords NER										✓										
German	CoreNLP	✓	✓		✓					✓											✓
	GATE Cloud - ANNIE	✓	✓		✓	✓	✓	✓	✓											✓	✓
	TagMe										✓										
	Keywords NER										✓										
French	CoreNLP	✓	✓			✓							✓								✓
	GATE Cloud - ANNIE	✓	✓		✓												✓	✓	✓	✓	✓
	Keywords NER										✓										
Spanish	CoreNLP	✓	✓		✓	✓				✓			✓								✓
	Keywords NER										✓										

Table I. Overview of the languages, annotations, and methods supported by the NLPHub. Check marks indicate if the annotation in the column is supported by the method and language in the row.

			Reference									
			Manual Annotation - I-CAB					AMERGE				
Annotation	Algorithm	Measure	Person	Geopolitical	Location	Organization	All	Person	Geopolitical	Location	Organization	Keyword
Person	**ItaliaNLP**	F-measure	79%	3%	0%	6%	51%	89%	3%	4%	10%	17%
		Precision	74%	2%	0%	5%	79%	100%	2%	4%	12%	54%
		Recall	84%	4%	1%	7%	38%	80%	5%	5%	8%	10%
	CoreNLP-Tint	F-measure	85%	2%	0%	4%	59%	95%	3%	3%	8%	25%
		Precision	78%	1%	0%	3%	81%	100%	2%	2%	9%	61%
		Recall	93%	4%	0%	6%	47%	91%	7%	4%	8%	16%
	AMERGE	F-measure	84%	3%	0%	7%	62%	100%	3%	4%	13%	26%
		Precision	74%	2%	0%	5%	80%	100%	2%	3%	13%	60%
		Recall	96%	6%	2%	10%	51%	100%	9%	7%	13%	17%
Geopolitical	**ItaliaNLP**	F-measure	1%	77%	4%	6%	29%	3%	100%	69%	11%	13%
		Precision	1%	74%	3%	8%	78%	8%	100%	83%	23%	81%
		Recall	1%	80%	7%	5%	18%	2%	100%	59%	7%	7%
	AMERGE	F-measure	0%	77%	4%	6%	29%	3%	100%	69%	11%	13%
		Precision	1%	74%	3%	8%	78%	8%	100%	83%	23%	81%
		Recall	1%	80%	7%	5%	18%	2%	100%	59%	7%	7%
Location	**ItaliaNLP**	F-measure	0%	1%	59%	1%	11%	2%	3%	41%	3%	4%
		Precision	2%	2%	52%	2%	50%	7%	4%	100%	10%	41%
		Recall	0%	1%	69%	1%	6%	1%	2%	26%	2%	2%
	CoreNLP-Tint	F-measure	1%	73%	30%	5%	40%	5%	77%	99%	13%	19%
		Precision	1%	63%	18%	5%	74%	8%	69%	100%	20%	76%
		Recall	1%	86%	84%	5%	27%	4%	87%	99%	10%	11%
	AMERGE	F-measure	1%	72%	31%	5%	39%	5%	76%	100%	13%	19%
		Precision	1%	62%	19%	5%	73%	8%	68%	100%	20%	75%
		Recall	1%	86%	88%	5%	27%	4%	87%	100%	10%	11%
Organization	**ItaliaNLP**	F-measure	3%	9%	3%	58%	35%	6%	10%	10%	79%	17%
		Precision	3%	7%	2%	52%	59%	8%	8%	10%	100%	61%
		Recall	3%	11%	7%	66%	25%	5%	14%	11%	65%	10%
	CoreNLP-Tint	F-measure	5%	9%	2%	65%	45%	12%	15%	12%	95%	25%
		Precision	4%	6%	1%	53%	59%	13%	10%	9%	100%	61%
		Recall	7%	16%	4%	83%	37%	12%	28%	16%	91%	16%
	AMERGE	F-measure	6%	11%	2%	63%	47%	14%	15%	14%	100%	28%
		Precision	5%	7%	1%	49%	57%	14%	10%	11%	100%	60%
		Recall	8%	21%	9%	87%	40%	14%	32%	21%	100%	18%
Keyword	**Keywords NER**	F-measure	20%	14%	6%	22%	40%	25%	17%	19%	32%	92%
		Precision	12%	8%	3%	13%	30%	17%	10%	11%	22%	99%
		Recall	56%	66%	58%	66%	61%	47%	66%	60%	56%	86%
	TagMe	F-measure	23%	33%	9%	25%	47%	23%	32%	31%	25%	41%
		Precision	18%	22%	5%	19%	55%	23%	21%	23%	26%	100%
		Recall	30%	67%	42%	38%	41%	23%	66%	47%	24%	26%
	AMERGE	F-measure	20%	18%	6%	22%	45%	26%	18%	21%	32%	100%
		Precision	12%	10%	3%	13%	32%	17%	10%	12%	21%	100%
		Recall	69%	91%	74%	79%	77%	60%	87%	77%	65%	100%

Table II. Performance calculation per annotation of all methods integrated with the NLPHub. Each cell reports the F-measure, precision, and recall of an annotator with respect to a reference annotation, alternatively given by the I-CAB corpus and the AMERGE algorithm. Using AMERGE as reference is useful to appreciate the overlap between the individual methods' annotations. The "All" annotation is a class containing all manually annotated named entities.

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

		Reference																		
		Manual Annotation - I-CAB					AMERGE					ItaliaNLP				CoreNLP-Tint			Keywords NER	TagMe
Annotation	Method	Person	Geopolitical	Location	Organization	All	Person	Geopolitical	Location	Organization	Keyword	Person	Geopolitical	Location	Organization	Person	Location	Organization	Keyword	Keyword
Person	**ItaliaNLP**	Excellent	Poor	Poor	Poor	Good	Excellent	Poor	Poor	Poor	Poor		Poor	Poor	Poor	Excellent	Poor	Poor	Poor	Marginal
	CoreNLP-Tint	Excellent	Poor	Poor	Poor	Good	Excellent	Poor	Poor	Poor	Poor	Excellent	Poor	Poor	Poor		Poor	Poor	Poor	Marginal
	AMERGE	Excellent	Poor	Poor	Poor	Good		Poor	Poor	Poor	Poor	Excellent	Poor	Poor	Poor	Excellent	Poor	Poor	Poor	Marginal
Geopolitical	**ItaliaNLP**	Poor	Good	Marginal	Poor	Marginal	Poor	Excellent	Good	Marginal	Marginal	Poor		Poor	Marginal	Poor	Good	Marginal	Poor	Marginal
	AMERGE	Poor	Good	Marginal	Poor	Marginal	Poor		Good	Marginal	Marginal	Poor	Excellent	Poor	Marginal	Poor	Good	Marginal	Marginal	Marginal
Location	**ItaliaNLP**	Poor	Poor	Good	Poor	Marginal	Poor	Poor	Marginal	Poor	Poor	Poor	Poor		Poor	Poor	Marginal	Poor	Poor	Marginal
	CoreNLP-Tint	Poor	Good	Marginal	Poor	Marginal	Poor	Good	Excellent	Marginal	Marginal	Poor	Good	Marginal	Marginal	Poor		Poor	Poor	Marginal
	AMERGE	Poor	Good	Marginal	Poor	Marginal	Poor	Good		Marginal	Marginal	Poor	Good	Marginal	Marginal	Poor	Excellent	Poor	Marginal	Marginal
Organization	**ItaliaNLP**	Poor	Marginal	Poor	Good	Marginal	Poor	Marginal	Marginal	Excellent	Poor	Poor	Marginal	Poor		Poor	Marginal	Good	Poor	Marginal
	CoreNLP-Tint	Poor	Marginal	Poor	Good	Marginal	Poor	Marginal	Poor	Excellent	Poor	Poor	Marginal	Poor	Good	Poor	Poor	Excellent	Poor	Marginal
	AMERGE	Poor	Marginal	Poor	Good	Marginal	Poor	Marginal	Marginal		Marginal	Poor	Marginal	Poor	Excellent	Poor	Marginal		Marginal	Marginal
Keyword	**Keywords NER**	Marginal	Marginal	Marginal	Marginal	Marginal	Poor	Marginal	Marginal	Marginal	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor		Marginal
	TagMe	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	Marginal	
	AMERGE	Marginal	Marginal	Marginal	Marginal	Marginal	Poor	Marginal	Marginal	Marginal		Poor	Marginal	Poor	Poor	Poor	Marginal	Poor	Excellent	Marginal

Table III. Agreement calculation as Fleiss' interpretation of Cohen's Kappa (overlap between two annotations with respect to co-annotation by chance). Each cell reports the agreement between the annotation/method in the column and the annotation/method in the row. Manual reference annotation is based on the I-CAB corpus.