AnnoMarket: An Open Cloud Platform for NLP

Valentin Tablan, Kalina Bontcheva
Ian Roberts, Hamish Cunningham
University of Sheffield,
Department of Computer Science
211 Portobello, Sheffield, UK
Initial.Surname@dcs.shef.ac.uk

Marin Dimitrov
Ontotext AD
47A Tsarigradsko Shosse, Sofia, Bulgaria
marin.dimitrov@ontotext.com

Abstract

This paper presents AnnoMarket, an open cloud-based platform which enables researchers to deploy, share, and use language processing components and resources, following the data-as-a-service and software-as-a-service paradigms. The focus is on multilingual text analysis resources and services, based on an open-source infrastructure and compliant with relevant NLP standards. We demonstrate how the AnnoMarket platform can be used to develop NLP applications with little or no programming, to index the results for enhanced browsing and search, and to evaluate performance. Utilising AnnoMarket is straightforward, since cloud infrastructural issues are dealt with by the platform, completely transparently to the user: load balancing, efficient data upload and storage, deployment on the virtual machines, security, and fault tolerance.

1 Introduction

Following the Software-as-a-Service (SaaS) paradigm from cloud computing (Dikaiakos et al., 2009), a number of text processing services have been developed, e.g. OpenCalais[1] and Alchemy API[2]. These provide information extraction services, accessible programmatically and charged per number of documents processed.

However, they suffer from two key technical drawbacks. Firstly, document-by-document processing over HTTP is inefficient on large datasets and is also limited to within-document text processing algorithms. Secondly, the text processing algorithms are pre-packaged: it is not possible for researchers to extend the functionality (e.g. adapt such a service to recognise new kinds of entities). Additionally, these text processing SaaS sites come with daily rate limits, in terms of number of API calls or documents that can be processed. Consequently, using these services for research is not just limited in terms of text processing functionality offered, but also quickly becomes very expensive on large-scale datasets. A moderately-sized collection of tweets, for example, comprises small but numerous documents, which can lead to unfeasibly high processing costs.

Platform-as-a-Service (PaaS) (Dikaiakos et al., 2009) are a type of cloud computing service which insulates developers from the low-level issues of utilising cloud infrastructures effectively, while providing facilities for efficient development, testing, and deployment of software over the Internet, following the SaaS model. In the context of traditional NLP research and development, and pre-dating cloud computing, similar needs were addressed through NLP infrastructures, such as GATE (Cunningham et al., 2013) and UIMA (Ferrucci and Lally, 2004). These infrastructures accelerated significantly the pace of NLP research, through reusable algorithms (e.g. rule-based pattern matching engines, machine learning algorithms), free tools for low-level NLP tasks, and support for multiple input and output document formats (e.g. XML, PDF, DOC, RDF, JSON).

This demonstration introduces the AnnoMarket[3] open, cloud-based platform, which has been developed following the PaaS paradigm. It enables researchers to deploy, share, and use language processing components and resources, following the Data-as-a-Service (DaaS) and Software-as-a-Service (SaaS) paradigms. It gives researchers access to an open, standard-compliant NLP infrastructure and enables them

[1] http://www.opencalais.com
[2] http://www.alchemyapi.com
[3] At the time of writing, a beta version of AnnoMarket is available at http://annomarket.com

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 19–24,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

to carry out large-scale NLP experiments by harnessing the vast, on-demand compute power of the Amazon cloud. It supports not only NLP algorithm development and execution, but also on-demand collaborative corpus annotation and performance evaluation. Important infrastructural issues are dealt with by the platform, completely transparently for the researcher: load balancing, efficient data upload and storage, deployment on the virtual machines, security, and fault tolerance.

AnnoMarket differs from previous work (e.g. (Zhou et al., 2010; Ramakrishnan et al., 2010)) in that it requires no programming in order to run a GATE-compliant NLP application on a large dataset. In that sense, it combines the ease of use of an NLP SaaS with the openness and comprehensive facilities of the GATE NLP infrastructure. AnnoMarket offers a growing number of pre-packaged services, in multiple languages. Additionally, as a specialised NLP PaaS, it also supports a *bring-your-own-pipeline* option, which can be built easily by reusing pre-existing GATE-compatible NLP components and adding some new ones. Moreover, in addition to offering entity extraction services like OpenCalais, our NLP PaaS also supports manual corpus annotation, semantic indexing and search, and performance evaluation.

The contributions of this paper are as follows:

1. A demonstration of running AnnoMarket multilingual NLP services on large datasets, without programming. The new service deployment facilities will also be shown, including how services can optionally be shared with others.
2. A demonstration on shared research corpora via the AnnoMarket platform, following the data-as-a-service model (the sharer is responsible for ensuring no copyright violations).
3. A demonstration of the large-scale search and browsing interface, which uses the results of the NLP SaaS to offer enhanced, semantic-based functionality.

2 The AnnoMarket NLP PaaS

This section first discusses the methodology underpinning the AnnoMarket platform, then presents its architecture and key components.

2.1 Development and Deployment Methodology

The development of text analysis algorithms and pipelines typically follows a certain methodological pattern, or lifecycle. A central problem is to define the NLP task, such that human annotators can perform it with a high level of agreement and to create high quality training and evaluation datasets. It is common to use double or triple annotation, where several people perform the annotation task independently and we then measure their level of agreement (*Inter-Annotator Agreement*, or IAA) to quantify and control the quality of this data (Hovy, 2010).

The AnnoMarket platform was therefore designed to offer full methodological support for all stages of the text analysis development lifecycle:

1. Create an initial prototype of the NLP pipeline, testing on a small document collection, using the desktop-based GATE user interface (Cunningham et al., 2002);
2. If required, collect a gold-standard corpus for evaluation and/or training, using the GATE Teamware collaborative corpus annotation service (Bontcheva et al., 2013), running in AnnoMarket;
3. Evaluate the performance of the automatic pipeline on the gold standard (either locally in the GATE development environment or on the cloud). Return to step 1 for further development and evaluation cycles, as needed.
4. Upload the large datasets and deploy the NLP pipeline on the AnnoMarket PaaS;
5. Run the large-scale NLP experiment and download the results as XML or a standard linguistic annotation format (Ide and Romary, 2004). AnnoMarket also offers scalable semantic indexing and search over the linguistic annotations and document content.
6. Analyse any errors, and if required, iterate again over the earlier steps.

AnnoMarket is fully compatible with the GATE open-source architecture (Cunningham et al., 2002), in order to benefit from GATE's numerous reusable and multilingual text processing components, and also from its infrastructural support for linguistic standards and diverse input formats.

2.2 Architecture

The architecture of the AnnoMarket PaaS comprises of four layers (see Figure 1), combining

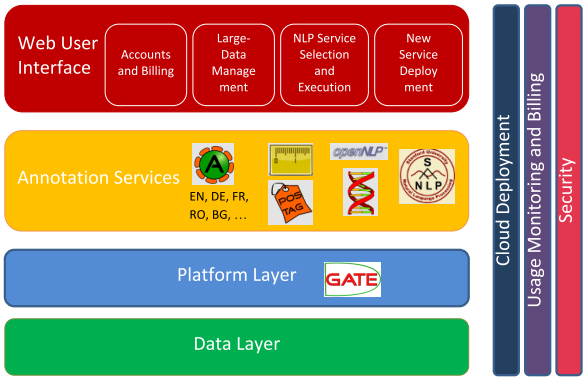

Figure 1: The AnnoMarket Architecture

components with related capabilities. Additionally, we have identified three aspects, which span across multiple layers.

The Data Layer is described in Section 2.3, the Platform Layer – in Section 2.4, and the Annotation Services – in Section 2.5.

The fourth, web user interface layer, contains a number of UI components that allow researchers to use the AnnoMarket platform in various ways, e.g. to run an already deployed text annotation service on a large dataset, to deploy and share a new service on the platform, or to upload (and optionally share) a document collection (i.e. a corpus). There is also support for finding relevant services, deployed on the AnnoMarket platform. Lastly, due to the platform running on the Amazon cloud infrastructure, there are account management interfaces, including billing information, payments, and usage reports.

The first vertical aspect is cloud deployment on Amazon. This covers support for automatic up and down-scaling of the allocated Amazon resources, detection of and recovery from Amazon infrastructure failures and network failures, and data backup.

Usage monitoring and billing is the second key vertical aspect, since fine-grained pay-as-you-go ability is essential. Even in the case of freely-available annotations services, Amazon usage charges are incurred and thus such functionality is needed. Various usage metrics are monitored and metered so that proper billing can be guaranteed, including: storage space required by language resources and data sets; CPU utilisation of the annotation services; number and size of documents processed.

Security aspects also have impact on all the layers of the AnnoMarket platform:

- Data Layer – data encryption and access control;
- Platform Layer – data encryption, authentication and access control;
- Service layer – authentication and transport level encryption;
- User Interface layer – authentication and transport level encryption.

In addition, we have implemented a REST programming API for AnnoMarket, so that data upload and download and running of annotation services can all be done automatically, outside of the web interface. This allows tighter integration within other applications, as well as support for synchronous (i.e. document-by-document) calling of the annotation services.

2.3 The Data Layer

The Data Layer stores various kinds of content, e.g. crawled web content, users' own corpora (private or shared with others), results from running the annotation services, etc.

Input documents can be in all major formats (e.g., XML, HTML, JSON, PDF, DOC), based on GATE's comprehensive format support. In all cases, when a document is being processed by AnnoMarket, the format is analysed and converted into a single unified, graph-based model of *annotation*: the one of the GATE NLP framework (Cunningham et al., 2002). Then this internal annotation format is also used by the collaborative corpus annotation web tool, and for annotation indexing and search. Annotations produced can be exported as in-line or stand-off XML, including XCES (Ide and Romary, 2004).

In implementation terms, Amazon S3 is used to store content on the platform. S3 provides a REST service for content access, as well as direct HTTP access, which provides an easy way for AnnoMarket users to upload and download content.

While stored on the cloud, data is protected by Amazon's security procedures. All transfers between the cloud storage, the annotation services, and the user's computer are done via an encrypted channel, using SSL.

2.4 The Platform Layer

The AnnoMarket platform provides an environment where text processing applications can be deployed as annotation services on the cloud. It allows processing pipelines that were produced on a

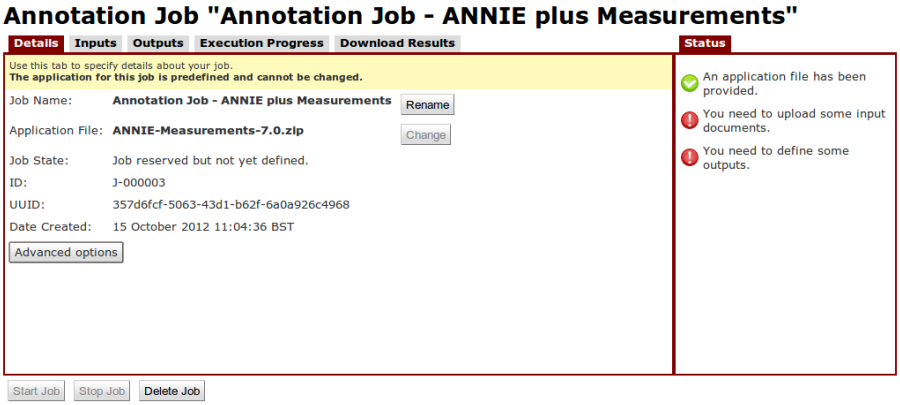

Figure 2: Web-based Job Editor

developer's stand-alone computer to be deployed seamlessly on distributed hardware resources (the compute cloud) with the aim of processing large amounts of data in a timely fashion. This process needs to be resilient in the face of failures at the level of the cloud infrastructure, the network communication, errors in the processing pipeline and in the input data.

The platform layer determines the optimal number of virtual machines for running a given NLP application, given the size of the document collection to be processed and taking into account the overhead in starting up new virtual machines on demand. The implementation is designed to be robust in the face of hardware failures and processing errors. For technical details on the way this was implemented on Amazon EC2 see (Tablan et al., 2013).

The GATE plugin-based architecture (Cunningham et al., 2002) is the basis for the platform environment. Users can upload any pipelines compliant with the GATE Processing Resource (PR) model and these are automatically deployed as annotation services on the AnnoMarket platform.

2.5 Annotation Services

As discussed above, the platform layer in AnnoMarket addresses most of the technical and methodological requirements towards the NLP PaaS, making the deployment, execution, and sharing of annotation services (i.e. pipelines and algorithms) a straightforward task. From a researcher's perspective, executing an annotation service on a dataset involves a few simple steps:

- Upload the document collection to be processed or point the system to a shared dataset on the platform;
- Upload a GATE-based processing pipeline to be used (or choose an already deployed annotation service);
- Set any required parameter values;
- Press the '*Start*' button.

While the job is running, a regularly updated execution log is made available in the user's dashboard. Upon job completion, an email notification is also sent. Most of the implementation details are hidden away from the user, who interacts with the system through a web-based job editor, depicted in Figure 2, or through a REST API.

The number of already deployed annotation services on the platform is growing continuously. Figure 3 shows a subset of them, as well as the metadata tags associated with these services, so that users can quickly restrict which types of services they are after and then be shown only the relevant subset. At the time of writing, there are services of the following kinds:

- Part-of-Speech-Taggers for English, German, Dutch, and Hungarian.
- Chunking: the GATE NP and VP chunkers and the OpenNLP ones;
- Parsing: currently the Stanford Parser [4], but more are under integration;
- Stemming in 15 languages, via the Snowball stemmer;
- Named Entity Recognition: in English, German, French, Arabic, Dutch, Romanian, and Bulgarian;
- Biomedical taggers: the PennBio[5] and the AbGene (Tanabe and Wilbur, 2002) taggers;
- Twitter-specific NLP: language detection, tokenisation, normalisation, POS tagging, and

[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] http://www.seas.upenn.edu/~strctlrn/BioTagger/BioTagger.html

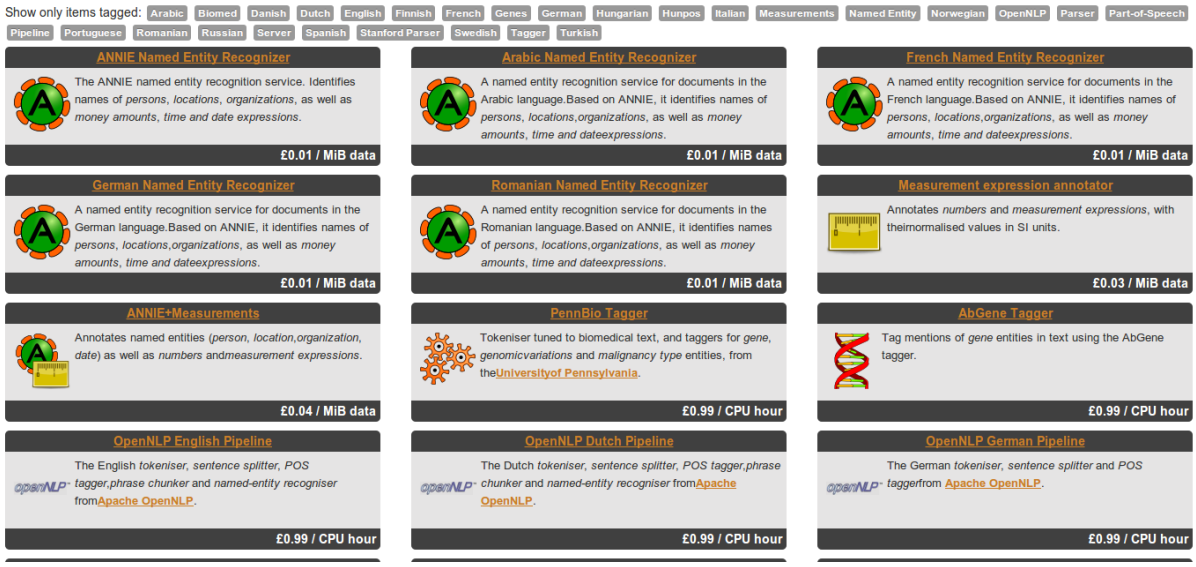

Figure 3: Pre-deployed Text Annotation Services

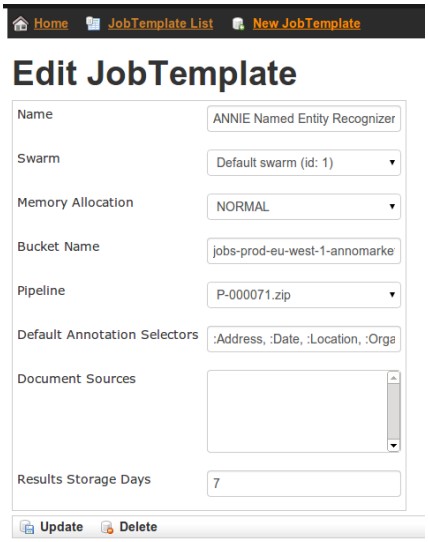

Figure 4: Creating a New Annotation Service

NER.

The deployment of new annotation services is done via a web interface (see Figure 4), where an administrator needs to configure some basic details related to the utilisation of the platform layer and provide a self-contained GATE-compatible application. Platform users can only publish their own annotation services by contacting an administrator, who can validate the provided pipeline before making it publicly available to the other users. This step is intended to protect the users community from malicious or poor quality pipelines.

3 Search and Browsing of Annotated Corpora

The AnnoMarket platform also includes a service for indexing and searching over a collection of semantically annotated documents. The output of an annotation service (see Figure 2) can be fed directly into a search index, which is created as the service is run on the documents. This provides facilities for searching over different views of document text, for example one can search the document's words, the part-of-speech of those words, or their morphological roots. As well as searching the document text, we also support searches over the documents' semantic annotations, e.g. named entity types or semantic roles.

Figure 5 shows a semantic search over 80,000 news web pages from the BBC. They have first been pre-processed with the POS tagging, morphological analysis, and NER services on the platform and the output indexed automatically. The search query is for documents, where entities of type Person are followed by any morphological form of the verb say, i.e. `{Person} root:say`.

4 Conclusion

This paper described a cloud-based open platform for text mining, which aims to assist the development and deployment of robust, large-scale text processing applications. By supporting the sharing of annotation pipelines, AnnoMarket also pro-

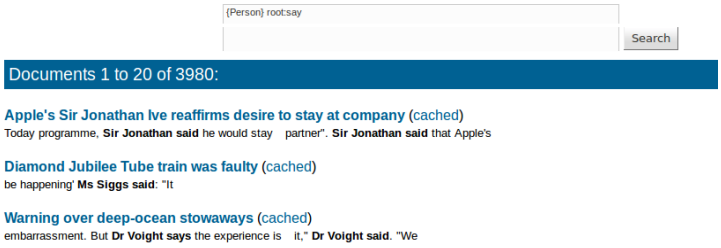

Figure 5: Example Semantic Search Results

motes reuse and repeatability of experiments.

As the number of annotation services offered by the platform has grown, we identified a need for service search, so that users can locate useful NLP services more effectively. We are currently developing a new UI, which offers search and browsing functionality, alongside various criteria, such as functionality (e.g. POS tagger, named entity recogniser), user ratings, natural language supported). In the medium- to long-term we have also planned to support UIMA-based pipelines, via GATE's UIMA compatibility layer.

A beta version is currently open to researchers for experimentation. Within the next six months we plan to to solicit more shared annotation pipelines to be deployed on the platform by other researchers.

Acknowledgments

This work was supported by the European Union under grant agreement No. 296322 AnnoMarket,[6] and a UK EPSRC grant No. EP/I004327/1.

References

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus. Roberts, Valentin. Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. GATE Teamware: A Web-based, Collaborative Text Annotation Framework. *Language Resources and Evaluation*.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. Gate: an architecture for development of robust hlt applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 7–12 July 2002*, ACL '02, pages 168–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, 02.

Marios D Dikaiakos, Dimitrios Katsaros, Pankaj Mehra, George Pallis, and Athena Vakali. 2009. Cloud computing: Distributed internet computing for IT and scientific research. *IEEE Internet Computing*, 13(5):10–13.

David Ferrucci and Adam Lally. 2004. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348.

Eduard Hovy. 2010. Annotation. In *Tutorial Abstracts of ACL*.

Nancy Ide and Laurent Romary. 2004. Standards for language resources. *Natural Language Engineering*, 10:211–225.

C. Ramakrishnan, W. A. Baumgartner, J. A. Blake, G. A. P. C. Burns, K. Bretonnel Cohen, H. Drabkin, J. Eppig, E. Hovy, C. N. Hsu, L. E. Hunter, T. Ingulfsen, H. R. Onda, S. Pokkunuri, E. Riloff, C. Roeder, and K. Verspoor. 2010. Building the scientific knowledge mine (SciKnowMine): a community-driven framework for text mining tools in direct service to biocuration. In *New Challenges for NLP Frameworks (NLPFrameworks 2010)*, LREC 2010, pages 9–14, Valletta, Malta, May. ELRA.

Valentin Tablan, Ian Roberts, Hamish Cunningham, and Kalina Bontcheva. 2013. GATECloud.net: a Platform for Large-Scale, Open-Source Text Processing on the Cloud. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, 371(1983):20120071.

Lorraine Tanabe and W. John Wilbur. 2002. Tagging Gene and Protein Names in Full Text Articles. In *Proceedings of the ACL-02 workshop on Natural Language Processing in the biomedical domain, 7–12 July 2002*, volume 3, pages 9–13, Philadelphia, PA. Association for Computational Linguistics.

Bin Zhou, Yan Jia, Chunyang Liu, and Xu Zhang. 2010. A distributed text mining system for online web textual data analysis. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2010 International Conference on*, pages 1–4, Los Alamitos, CA, USA, October. IEEE Computer Society.

[6] See `http://www.annomarket.eu/`.