AnnoMarket – Multilingual Text Analytics at Scale on the Cloud

Marin Dimitrov[1], Hamish Cunningham[2], Ian Roberts[2], Petar Kostov[1], Alex Simov[1], Philippe Rigaux[3], Helen Lippell[4]

[1] Ontotext AD, Bulgaria
{marin.dimitrov, petar.kostov, alex.simov}@ontotext.com
[2] Department of Computer Science, University of Sheffield, UK
{h.cunningham, i.roberts}@ dcs.shef.ac.uk
[3] Internet Memory Research SAS, France
philippe.rigaux@internetmemory.net
[4] The Press Association Ltd, UK
helen.lippell@pressassociation.com

Abstract. AnnoMarket is an open platform for cloud-based text analytics services and language resources acquisition. Providers of text analytics services and language resources can deploy and monetize their components via the platform, while users can utilize such available resources in multiple languages and in various domains in an on-demand, pay-as-you-go manner. The AnnoMarket platform is deployed on the Amazon Web Services cloud and it provides free text analytics and language acquisition services to the general public.

Keywords: text mining, cloud computing, software-as-a-service, linked data

1 Introduction

AnnoMarket[1] is an FP7[2] project that aims to revolutionise the text analytics market, by delivering an open marketplace for pay-as-you-go, cloud-based text mining resources and services, in multiple languages. The current services available on the AnnoMarket marketplace[3] are applicable to a wide set of business cases, e.g. large-volume multilingual information management, business intelligence, social media monitoring, customer relations management.

The Software-as-a-Service delivery model adopted by AnnoMarket reduces the complexity of deployment, maintenance, customisation and sharing of text processing resources and services by SMEs and developers. Additional beneficiaries of AnnoMarket are SME providers of text analytics services or language resources, who are able to deploy their custom components, applications, datasets or corpora, and receive

1 https://annomarket.eu/

2 The AnnoMarket project is funded by the European Commission under the 7th Framework Programme, Project No. 296322

3 https://annomarket.com/

adfa, p. 1, 2011.

revenue via the AnnoMarket marketplace. The marketplace currently provides various services for multilingual information extraction and semantic annotation, sentiment detection, as well as multilingual corpora and LOD datasets.

2 AnnoMarket Platform

The AnnoMarket platform is based on the *GATE* [1] and *GateCloud.net* [2] platforms with various new components related to language resource acquisition, scalable and elastic processing of large volumes of data, usage monitoring and quota enforcement, as well as billing and online payments.

2.1 Language Resource Acquisition

The language resource acquisition component of the platform is based on the large scale web crawling infrastructure by IMR comprised of three main components:

- *MemoryBot*, a scalable web crawler which provides user-defined, on-demand crawls at a large scale
- An integrated annotation mechanism, which provides means for pre-processing of the crawled corpora (feature extraction, statistical information, etc.) and generates various metadata that is utilized for indexing and searching
- A distributed language resource repository, based on HBase, which stores the original crawled content and the metadata from the pre-processing step.

In addition to corpora crawled on-demand from the web, the language resource acquisition component provides integration with the Common Crawl[4] dataset.

2.2 Multilingual Text Mining Services

Various multilingual text mining services are currently deployed and ready to use via the AnnoMarket platform. The current set of services includes more than 30 different text processing pipelines covering 17 languages: Arabic, Bulgarian, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish, and Turkish.

The text processing pipelines vary from low level ones (stemmers, part-of-speech taggers, noun phrase chunkers and parsers), to general purpose pipelines (named entity recognisers) and domain specific pipelines (for the bio-medical domain, news publishing domain, or sentiment analysis over social media).

2.3 Marketplace

The marketplace provides an eShop[5] where customers can explore the catalogue of available text analytics services, language resources and datasets as well as additional

[4] http://commoncrawl.org/

processing resources available on-demand via the platform (e.g. an LOD server hosting Freebase, DBpedia and GeoNames datasets which can be used to populate various gazetteers for text mining pipelines). All products deployed on the marketplace provide information about their functionality and the associated usage and pricing terms. Text analytics services may also show a simple web form where customers can supply sample input data and test the functionality of the service. Customers can also post public ratings and comments regarding the performance and quality of service of any product they have used via the AnnoMarket platform[6].

2.4 Cloud Platform

The AnnoMarket platform is currently deployed on the Amazon Web Services[7] (AWS) public cloud and it utilizes various cloud services for storage (S3, EBS and SimpleDB), computing (EC2), and scalability (Simple Queue Service, Auto Scaling and CloudWatch), and a design for a multi-datacenter deployment for improved availability.

AnnoMarket customers can utilize the various services on the platform via two delivery channels:

- platform-as-a-service, where customers configure, start and stop various processing and storage components on demand and customers are billed for the duration of using the components (per hour);
- software-as-a-service (**Fig. 1**), where multiple customers can access the text analytics components in a multi-tenant manner via predefined RESTful service interfaces. Customers in this case are billed based on usage metrics (number of documents processed, input data size, number of SPARQL queries, etc.)

2.5 Workflow

The interaction with the AnnoMarket platform when using text analytics components in a platform-as-service manner includes the following steps:

1. Users specify the input data to be processed. This may include user data already residing on Amazon S3, language resources and datasets available on the platform, or a new on-demand focused crawl from the web.
2. Users choose and configure various processing and indexing services, which are automatically deployed and started on Amazon EC2, or access the RESTful services in a software-as-a-service manner.
3. After the processing is complete, the results are available to the customer via S3 and can be downloaded at any time.

---

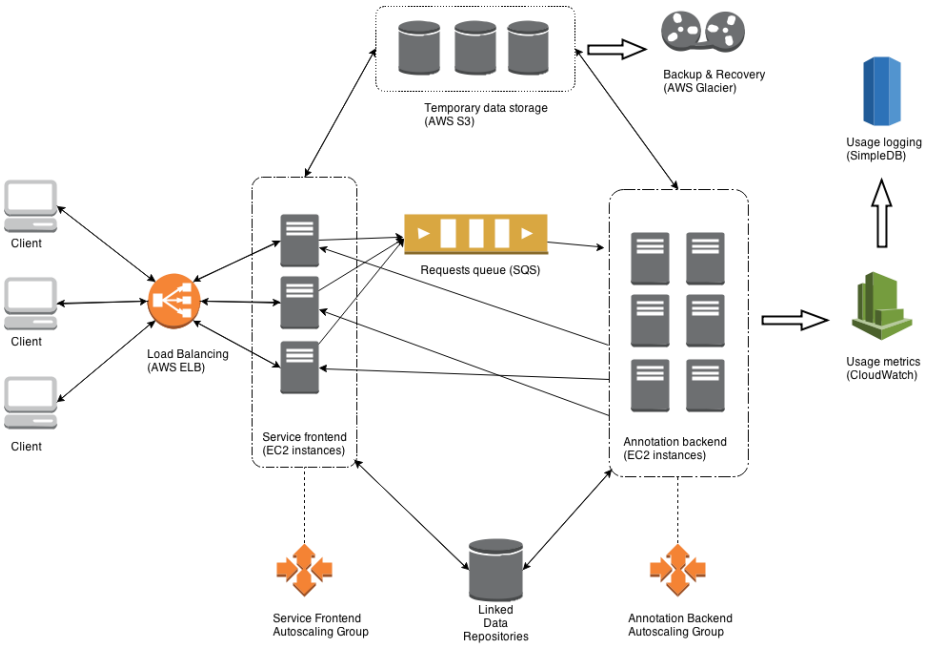
5 https://annomarket.com/shopfront

6 See for example https://annomarket.com/shopfront/displayItem/2 information available for the particular pipeline: developer's documentation and instructions, a "Test This Pipeline" component and a user comments section.

7 https://aws.amazon.com/

When using the text analytics components of the platform in a software-as-service manner users directly submit documents to the RESTful services, together with various specifications (requested output format, processing pipeline, etc.). The results of the processing are immediately available as part of the service response and do not require additional downloading.

Fig. 1. AnnoMarket AWS-based architecture



2.6 Client APIs and Browser Plugins

For the purpose of making first interactions with the AnnoMarket platform easier, we have developed a Client API and a browser plugin for Firefox and Chrome.

The Client API[8] provides a set of high-level APIs in Java, C#, Groovy and Python, which make interaction with the platform easier for developers (as opposed to directly making RESTful invocations).

3 AnnoMarket Demo

The AnnoMarket team will demonstrate the platform via various scenarios targeting different users:

8 https://github.com/annomarket/online-clients

1. For developers – a demonstration on (1) how to configure the MemoryBot crawler and to perform focussed crawling over a large number of websites; (2) how to access the real-time text analytics service via the Client API or directly via the RESTful service interface; and (3) how to configure a long-running text analytics jobs, provide input data, and run various text processing or indexing components.
2. For text analytics providers – how to package their own text analytics pipelines, so that they can be deployed on the AnnoMarket platform and available for use to 3rd parties via the Marketplace.

The AnnoMarket platform is currently open to the general public and its functionality can be accessed via:

- The marketplace[9], providing entry points to the various text analytics services deployed on the platform
- The real-time text processing RESTful services[10]
- The RESTful services[11] for configuring and starting long-running text annotation jobs
- The Client APIs[12] in various programming languages

4 Related Work

There exist various platforms for text analytics as-a-service, such as: *OpenCalais*, *Alechemy*, *OpenAmplify*, *Semantria*, *TextWise*, *Saplo*, etc., or on-demand crawling such as *80Legs*, *Spinn3R* and *PromptCloud*. *Mashape* is a somewhat similar platform and marketplace for 3rd party APIs. The main differentiation of AnnoMarket is a combination of the following aspects: 1) extensive list of text analytics components covering 17 languages; 2) ability of 3rd party text analytics providers to deploy and monetize their components via the platform; 3) a combination of on-demand crawling and text analytics capabilities.

The AnnoMarket platform is currently undergoing an evaluation by a 3rd party focus group of various performance, usability, and availability aspects.

References

1. H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. In *PLoS Computational Biology* 9(2). 2013
2. V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva. GATECloud.net: a platform for large-scale, open-source text processing on the cloud. In *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences*, vol. 371 no. 1983, 2012

[9] https://annomarket.com/shopfront
[10] https://api.annomarket.com/online-processing/item/<pipeline-number>
[11] https://annomarket.com/api/shop/item/<pipeline-number>
[12] https://github.com/annomarket/online-clients