xLiMe
FP7-ICT-2013-10
ICT-2013.4.1
Contract no.: 611346
www.xlime.eu

Deliverable D3.3.1

Early Prototype of Text Annotation

Editor:	Aditya Mogadala, KIT
Author(s):	Aditya Mogadala, KIT; Andreas Wagner, KIT
Deliverable Nature:	Prototype (P)
Dissemination Level:	Public (PU)
Contractual Delivery Date:	M12 - 31 October 2014
Actual Delivery Date:	M12 - 31 October 2014
Suggested Readers:	Researchers and developers who are interested in providing cross-language annotations using DBpedia on natural language text.
Version:	2.0
Keywords:	Semantic annotation, cross-language, knowledge bases

Disclaimer

Full Project Title:	Cross-lingual Cross-Media Knowledge Extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP3 Cross-lingual Multimedia Semantic Annotation
Document Title:	D3.3.1 – Early Prototype for Text Annotation
Editor:	Aditya Mogadala, KIT
Work Package Leader:	Nicu Sebe, UNITN

Executive Summary

The goal of this deliverable is to provide details of the light-weight text annotation approach built on the work accomplished in the XLike [1] project. The annotations identified are the links between words or phrases found in a document with the entities present in the knowledge bases like DBpedia. These annotations further help to facilitate monolingual and cross-lingual semantic search.

The documents on which annotations are performed vary in length and are collected from myriad sources. Also, the text present in these documents can be noisy as found in Tweets or linguistically rich as observed in News articles.

T3.3.1 deliverable also aims to cover three languages mainly German, English and Spanish for cross-lingual and mono-lingual annotations.

Table of Contents

Abbreviations

RDF	Resource Description Framework
URL	Unified Resource Locator
URI	Uniform Resource Identifier
VICO	VICO Research & Consulting GmbH
JSI	Jozef Stefan Institute

1 Introduction

In this document, we provide a detailed description of the early prototype developed for text annotation. The prototype is used for linking unstructured multi-lingual texts with the world knowledge present in knowledge bases. The contribution of the prototype is not only limited to mono-lingual text annotations in the documents. But, also to provide cross-lingual annotations by leveraging knowledge base cross-language links.

Let's consider an example. If an English document contains the word "Germany", we annotate the word with English DBpedia [**2**] entity "Germany" and identify cross-language annotations in German and Spanish with the DBpedia property "**owl:sameAs**".

The table below shows the DBpedia links for the entity.

Example Cross-Language Links in DBpedia

English DBpedia Link : http://dbpedia.org/page/Germany
German DBpedia Link: http://de.dbpedia.org/page/Deutschland
Spanish DBpedia Link: http://es.dbpedia.org/page/Alemania

Also as a pre-requisite for semantic search, the prototype produces the output satisfying the requirements of xLiMe Meta data model [**3**]. The xLiMe Meta data model is designed to store the data in RDF format. This facilitates both mono-lingual and cross-lingual semantic search by using different graph based similarity measures.

Using this prototype annotation can be performed on any form of text. The text can belong to blogs, review sites, forums, Facebook posts and Tweets. In the next section, we provide an example for annotations obtained on the VICO [**4**] social media text produced in xLiMe pipeline.

1.1 VICO Social Media Text Annotation

Social media annotations are part of xLiMe pipeline. The data generated from social media resources are first embedded into an xLiMe Meta data model along with their text annotations. The example below depicts the text annotations generated from a social media forum.

Example – VICO Social Media Text Annotation

```
@prefix sioc:  <http://rdfs.org/sioc/ns#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix kdo:   <http://kdo.render-project-eu/kdo#> .
@prefix prov:  <http://www.w3.org/ns/prov#> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix xlime: <http://xlime-project.org/vocab/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

{ <http://vico-research.com/social/bd5eff57-8419-44f2-8d84-adae742a8abc>
          dcterms:title          "KIT Annotations" ;
          prov:wasAttributedTo   <http://aifb.kit.edu> ;
          prov:wasGeneratedBy    [ a                    prov:Activity ;
                                   prov:endedAtTime     "2014-09-
01T14:36:55"^^xsd:dateTime ;
                                   prov:startedAtTime   "2014-09-
01T14:36:36"^^xsd:dateTime
```



```
                                ] .
}
<http://vico-
research.com/social/http://www.landtreff.de/index.php/2cbb23a7-b6ff-380f-
8803-efa9de52eef7>
            a                    sioc:MicroPost ;
            dcterms:created      "2014-08-25T21:51:00"^^xsd:dateTime ;
            dcterms:language     "de" ;
            dcterms:publisher    <http://www.landtreff.de/index.php> ;
            dcterms:source
<http://www.landtreff.de/post1197936.html#p1197936#5> ;
            dcterms:spatial      [ rdfs:label  "de" ] ;
            sioc:content         "Adidas und Puma" ;
            sioc:has_creator
<http://www.landtreff.de/index.php#WaldbauerSchosi> ;
            xlime:hasAnnotation  [ xlime:hasConfidence
"0.86"^^xsd:double ;
                                   xlime:hasEntity
<http://de.dbpedia.org/resource/Puma_%28Unternehmen%29> ;
                                   xlime:hasPosition    [
xlime:hasStartPosition  "11"^^xsd:long ;

xlime:hasStopPosition   "15"^^xsd:long
                                                        ]
                                 ] ;
            xlime:hasAnnotation  [ xlime:hasConfidence  "1"^^xsd:double ;
                                   xlime:hasEntity
<http://de.dbpedia.org/resource/Adidas> ;
                                   xlime:hasPosition    [
xlime:hasStartPosition  "0"^^xsd:long ;

xlime:hasStopPosition   "6"^^xsd:long
                                                        ]
                                 ] .
```

2 Approach for Text Annotation

In this section we first evaluate the various options available for knowledge bases and then describe our approach.

2.1 Knowledge Bases

The two widely used knowledge bases are Wikipedia and DBpedia. Each of them has different properties and can be used for various tasks. In the following sections we describe them briefly.

2.1.1 Wikipedia as Knowledge Base

Generally, Wikipedia is considered as an online encyclopaedia containing unstructured text with rich interlinked information. It consists of different concepts and semantic relations. Most of the text written in different languages about various concepts can be considered as a comparable corpus. This provides researchers and developers working with natural language text and semantic web a highly available resource for various tasks.

The content present in the Wikipedia articles consists of text descriptions, category information and hyperlinks to other articles. Each article can be considered as a concept, and each concept can be known using the Wikipedia article title. This information is very helpful in creating ontologies and thesauri.

Most of the Wikipedia articles provide terminology that associate different articles having similar surface forms. For Example, consider two different articles taking about "*Scandinavian unitary*" and "*land of mid night sun*". Even though both of these articles refer "*Norway*", it is really difficult to identify this information without any proper knowledge structure. Surprisingly, Wikipedia links both of these articles to a single article that describes "*Norway*".

The articles of Wikipedia whose titles share synonyms encode polysemy. For example, "jaguar" can be a large cat or a product from a car company. This phenomenon is observed in most articles that are referred from different links. This information is useful for providing disambiguation during annotation of semantic labels.

2.1.2 DBpedia as Knowledge Base

Most of the Wikipedia articles are semi-structured. It contains unstructured text and structured information present in info-boxes etc. DBpedia is an effort to build a structured knowledge base by extracting information from Wikipedia. English version 3.8 of DBpedia contains around 3.7 million entities. Out of them 2.35 million has been classified into proper ontology. There are around 764 thousand persons, 573 thousand places, 192 thousand organizations, 202 thousand species, 112 thousand music albums, 72,000 movies and 5500 diseases.

DBpedia is also present in different languages. Currently, it has around 110+ languages. There are around 20.8 million entities in these languages and most of them overlap with English DBpedia. Also, there are around 8 million links to images, 24.4 million links to external web pages, 27.2 million links to RDF data, 55.5 million to Wikipedia categories and 8.2 million YAGO categories. YAGO is a huge semantic knowledge base derived from Wikipedia, WordNet and GeoNames.

The dataset also consists of around 1.89 billion RDF triples. Out of which 403 million are extracted from the English Wikipedia and around 1.46 billion were extracted from other language collections. There are also 27 million links that refer to other RDF datasets.

2.2 Cross-lingual links in DBpedia

Entities of DBpedia are present in multiple languages. DBpedia lookup application provides every entity with an "**owl: sameAs**" property along with a value describing links to the cross-language peers in other languages. This information facilitates the cross-language annotations on the multi-lingual text.

Below, we provide a sample annotation obtained on the social media text with cross-language links using the data provided in the previous section. The example shows a German DBpedia annotation on the German text along with the cross-language links obtained from English and Spanish DBpedia. We can also see the confidence at which the annotations were performed using "xlime: hasConfidence" attribute.

```
<http://www.landtreff.de/index.php#WaldbauerSchosi>;
xlime:hasAnnotation [
                      xlime:hasConfidence  "0.86"^^xsd:double;
                      xlime:hasEntity
<http://de.dbpedia.org/resource/Puma_%28Unternehmen%29>;
                      xlime:hasPosition [
                            xlime:hasStartPosition  "11"^^xsd:long ;
                            xlime:hasStopPosition   "15"^^xsd:long ;
                    ]
<http://dbpedia.org/page/Puma_SE>;
<http://es.dbpedia.org/page/Puma_%28marca%29>;

                     ];
xlime:hasAnnotation [
                        xlime:hasConfidence  "1"^^xsd:double ;
                        xlime:hasEntity
<http://de.dbpedia.org/resource/Adidas> ;
                        xlime:hasPosition [
                                xlime:hasStartPosition  "0"^^xsd:long ;
                                xlime:hasStopPosition   "6"^^xsd:long
                         ]
<http://dbpedia.org/page/Adidas>;
<http://es.dbpedia.org/page/Adidas>;
                       ];
```

2.3 Text Annotation

In this section, we provide our approach to annotate the unstructured multi-lingual text with knowledge base entities. We choose DBpedia over Wikipedia to perform this annotation, due to its support for various applications like lookup and cross-language analysis.

Through DBpedia lookup service, DBpedia entities are identified for the key phrases or keywords found in a document. But, most of these words can be annotated with more than one DBpedia entity. To solve the problem of disambiguation, we adopt a four step approach similar to XLike [1] for ranking the entities during annotation. The four steps used for annotation are

(1) Extraction of Phrases or words (modified for social media)

(2) Selecting an appropriate DBpedia Entity,

(3) Graph Construction, and

(4) Entity Disambiguation.

Following sections provide the brief descriptions for each of these steps.

2.3.1 Extraction of Phrases or Words

Unigrams represent only words present inside a document. To find the phrases, we extract N-grams that are linguistically correct. In-order to identify the linguistically correct N-grams, we find their corresponding matches with the DBpedia entities.

In xLiMe as compared to XLike, we also deal with noisy text generated from social media forums like Twitter. Identifying phrases from the noisy text is a challenging task and can result in low recall of the identified entities. To overcome this problem, we cleaned the text using simple text normalization approaches.

Junk characters are eliminated and tweets containing languages other than English and German are discarded. Also, user id information is eliminated from the text of the tweet to extract phrases or keywords. If there is a match, we treat them as "*mentions*" of world knowledge in the text. But, there can be more than one "*mention*" for an N-gram. To find the right "*mention*", we need to rank the entities based on certain similarity measure. We explore the similarity measure in the following section.

2.3.2 Selecting a DBpedia Entity

Identification of the best possible DBpedia entity that can be used for annotation of an N-gram is obtained using a similarity metric. The metric ranks all possible DBpedia entities. The Similarity metric is a weighted combination of prior links of N-grams found in the DBpedia entities and the context in which N-grams were present in the document. Equation-1 represents the similarity metric used for annotation.

$$SM(nq, en) = \alpha.PL(nq, en) + \beta.SC(nq, en)$$

Equation-1

The SM (nq, en) represents the similarity measure between N-gram **(nq**) and the DBpedia entity (**en**). While, PL(nq,en) represents the prior link probability between **nq** and **en**. Similarly, SC (nq, en) represents the context similarity between **nq** and **en**. Alpha and Beta parameters are selected empirically to maximize the similarity.

PL (nq, en) is calculated by taking the ratio of possible entity matches for a given N-gram with the total entities of DBpedia given by ∑ en. Matching (nq, en) is calculated as count of sub-string matches for a given N-gram with the DBpedia entities. Equation-2 shows the representation of PL.

$$PL(nq, en) = \frac{Matching(nq, en)}{\sum en}$$

Equation-2

To find the context similarity of an N-gram and DBpedia entity, we find the likelihood of generating the entity in some context "C" where the N-gram is observed. This is calculated as a conditional probability where the context "C" is given by observing previous N-grams that had entity matches. Equation-3 represents SC (nq, en).

$$SC(nq, en) = P(C|en) = \prod_{s \epsilon nq, t \epsilon en} PL(s, t)$$

Equation-3

Where P (C|en) shows the conditional probability of finding the entity in context of "C" where an N-gram was found. By using the probability chain rule, we expand this conditional probability as the product of different N-grams found in a document in sequence along with entities matched.

2.3.3 Graph Construction

Once the list of probable DBpedia entities are obtained for all the extracted N-grams, we disambiguate them. Let's call the list of N-Grams as **NG** and DBpedia entities as **EN**. We form a set using the list **NG** and represent it using the notation [NG] and the list **EN** as another set represented as [EN]. We add the two sets into a graph **G** represented as [G].

If any candidate of the set [NG] is linked to any candidate of the set [EN], an edge is added between them. Further, for each pair of [EN], if there is a property between two entities in DBpedia. An edge is added between them.

Once the graph construction is done between N-grams and DBpedia entities, we find the semantic relatedness between them using a metric inspired from Normalized Google Distance (NGD). Metric is visualized in Equation-4.

$$SMR(e_i, e_j) = 1 - \frac{log(\frac{max(|E_i|,|E_j|)}{E_i \bigcap E_j})}{log(\frac{|M|}{min(|E_i|,|E_j|)})}$$

Equation-4

SMR represents the semantic relatedness between two different nodes represented as entities in the graph G. |M| represent the total number of entities present in DBpedia. While $|E_i|$and $|E_j|$ represents the size of either N-grams or DBpedia entities sets to which e_i and e_j belongs. For example, if e_i belongs to [NG] then E_i represents [NG]. If e_i belongs to [EN] then E_i represents [EN]. Similarly, it is interchanged for the e_j and E_j.

2.3.4 Entity Disambiguation

Once the graph G is built, the entity disambiguation is achieved with personalized PageRank (PPr) [**5**]. Final ranked list of entities for each N-Gram is obtained by applying PPr on the graph G. The entity on top of the list is considered as the best possible match for every N-gram. Let **M** denotes the mentions for the entities **E** found in DBpedia for the set of all N-grams.

PPr is a combination of voting scheme (VS) and random jumps (RJ). VS represent the product of prior PPr score of each page the random walker would visit and the new score it attains by traversing the new nodes in graph G. RJ provides the score of random jumps of walker when it encounters any dead links or spider trap in the graph G. Equation-5 shows the joint representation of VS and RJ score.

(VS) (RJ)

$$PPr = \alpha.T.PPr + (1-\alpha).v$$

Equation-5

Initial score of **T** is calculated using the links between the entities **E** and mentions **M** in graph G. Value of parameter **alpha** is assigned by empirical analysis. The value for **T** is calculated using the equations listed below.

$$T_{ij} = \frac{SMR(e_i, e_j)}{\sum_{k \epsilon G} SMR(e_i, e_k)} \quad \text{Where } i, j \epsilon E, i \rightarrow j$$

Equation-6

$$T_{ij} = \frac{Sim(e_i, e_j)}{\sum_{k \epsilon E} Sim(e_i, e_k)} \quad \text{Where } i \epsilon M, j \epsilon E, i \rightarrow j$$

Equation-7

T is assigned to **zero**, if does not satisfy any of the conditions listed in Equation-6 and Equation-7. Sim(e_i, e_j) represents the edge distance between the two entities found in their respective sets of **M** and **E** of the entire graph G.

Now we calculate the value **V** for the random jumps (RJ) using Equation-8 listed below.

$$v_i = \begin{cases} \frac{1}{M} & i \epsilon M \\ 0 & Otherwise \end{cases}$$

Equation-8

2.3.5 Linking other Datasets

The approach mentioned in the earlier sections is used to obtain DBpedia entities as mentions for all the N-grams found in a document. But, there are many resources like DBpedia that contains an entity or concept information. We find all those resources using DBpedia property for an entity called "**owl: sameAs**".

The value of this property contains links to entities of DBpedia present in different languages and also links to other knowledge bases.

3 Annotation Service

An annotation service is hosted to annotate the documents generated in the xLiMe pipeline.

3.1 Configuration

Below, we list down the input arguments and the expected output from our annotation service.

Input Type	**Arguments**
Source	The URL of a web page or plain text
Model	Model is used for identifying the mentions. Here, we use only N-Grams to identify mentions. **Example:** NGRAM
Source Language	Language to be used for the annotations and source document. **Example:** en,de,es
Target Language	Cross-language annotations for a source language document are obtained by mentioning the target language. **Example:** en,de,es
Knowledge Base	The knowledge base we want to use for annotation. **Example:** DBpedia

The output is generated in the XML format with annotations to DBpedia entities.

3.2 Sample Output

In this section, we present the sample output generated by our annotation service.

In this example, the input to the annotation service is a snippet of blog article generated by the JSI News content provider in the xLiMe pipeline. The output obtained after annotation is the links to entities in DBpedia same as the language of document.

The tables below show the input and the output generated by the annotation service. Please note that the “detected topics” are the “entities” of DBpedia.

INPUT – Plain Text Extracted from the RDF format of JSI NEWS stream

```
The threat from IS is largely there because of our 2003 invasion of
Iraq.I cannot think of one single military intervention from the west
into the Mid east that has actually worked out well .I question War
because it achieves very little apart from bringing terrorism to our own
shores
You may well be right but we can't change any of that and ignoring this
problem is only going to lead to things getting worse IMHO. These people
are not going to stop and if they're allowed to develop, they will start
to attack our interests directly and indirectly to the point where we
will have to fight back.
```

OUTPUT – With Annotations of DBpedia entities in XML Format.

```
<?xml version="1.0" encoding="UTF-8"?>
<AnnotationResponse>
  <Result>The threat from IS is largely there because of our [[2003
invasion of Iraq|intervention in Iraq]] .I cannot think of one single
[[military]] intervention from the west into the Mid east that has
actually worked out well .I question [[War|military action]] because it
achieves very little apart from bringing [[terrorism]] to our own shores
You may well be right but we can't change any of that and ignoring this
problem is only going to lead to things getting worse IMHO. These people
are not going to stop and if they're allowed to develop, they will start
to attack our interests directly and indirectly to the point where we
will have to fight back.</Result>
  <DetectedTopics>
    <DetectedTopic
URL="http://dbpedia.org/resource/2003_invasion_of_Iraq" displayName="2003
invasion of Iraq" id="201936" lang="en" mention="intervention in Iraq"
weight="0.693"/>
    <DetectedTopic URL="http://dbpedia.org/resource/Iraq"
displayName="Iraq" id="7515928" lang="en" mention="Iraq" weight="1"/>
    <DetectedTopic URL="http://dbpedia.org/resource/Military"
displayName="Military" id="92357" lang="en" mention="military"
weight="0.169"/>
    <DetectedTopic URL="http://dbpedia.org/resource/War"
displayName="War" id="33158" lang="en" mention="military action"
weight="0.181"/>
    <DetectedTopic URL="http://dbpedia.org/resource/Terrorism"
displayName="Terrorism" id="30636" lang="en" mention="terrorism"
weight="0.458"/>
  </DetectedTopics>
</AnnotationResponse>
```

3.3 VICO Usage

VICO is one of the use case partners who annotate social media streams to monitor and track entities for various business applications. In this section, we see a sample entity identified in DBpedia same as the language of the VICO social media stream.

Figure 1: example entity - “adidas”Figure 1 shows an example entity - “adidas” identified in the twitter stream.

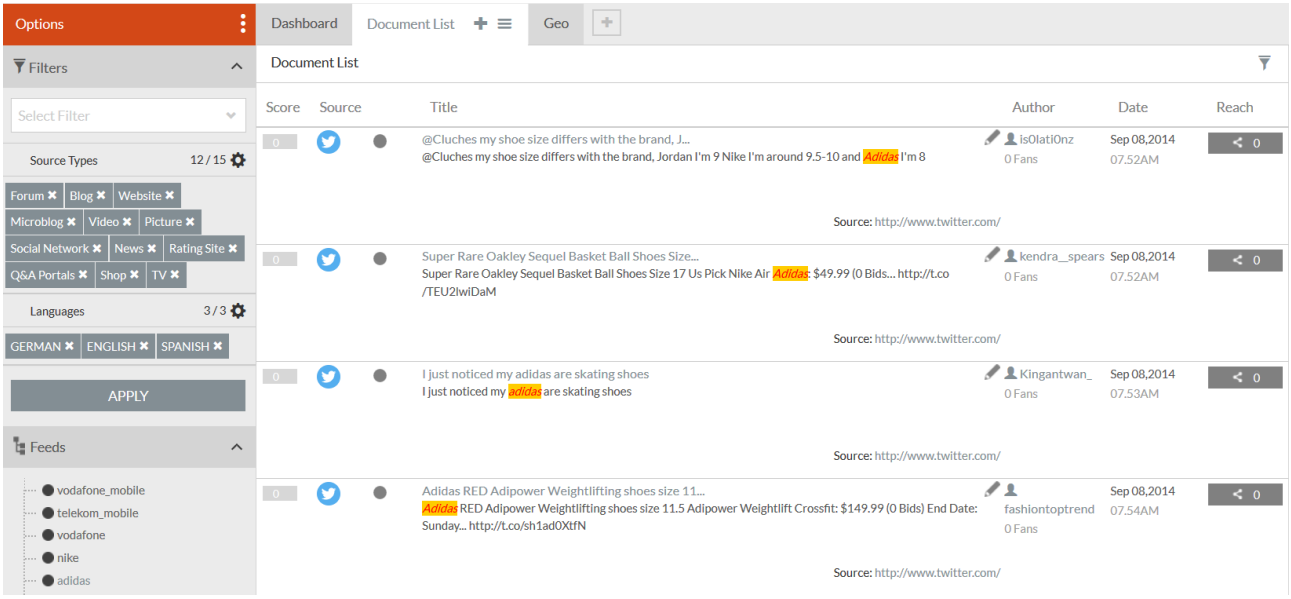

Figure 1: example entity - “adidas”

Similarly, we see the volume at which the entity “adidas” was found in the social media stream. This helps to monitor and track entities over time. Figure 2 visualizes the entity tracking.

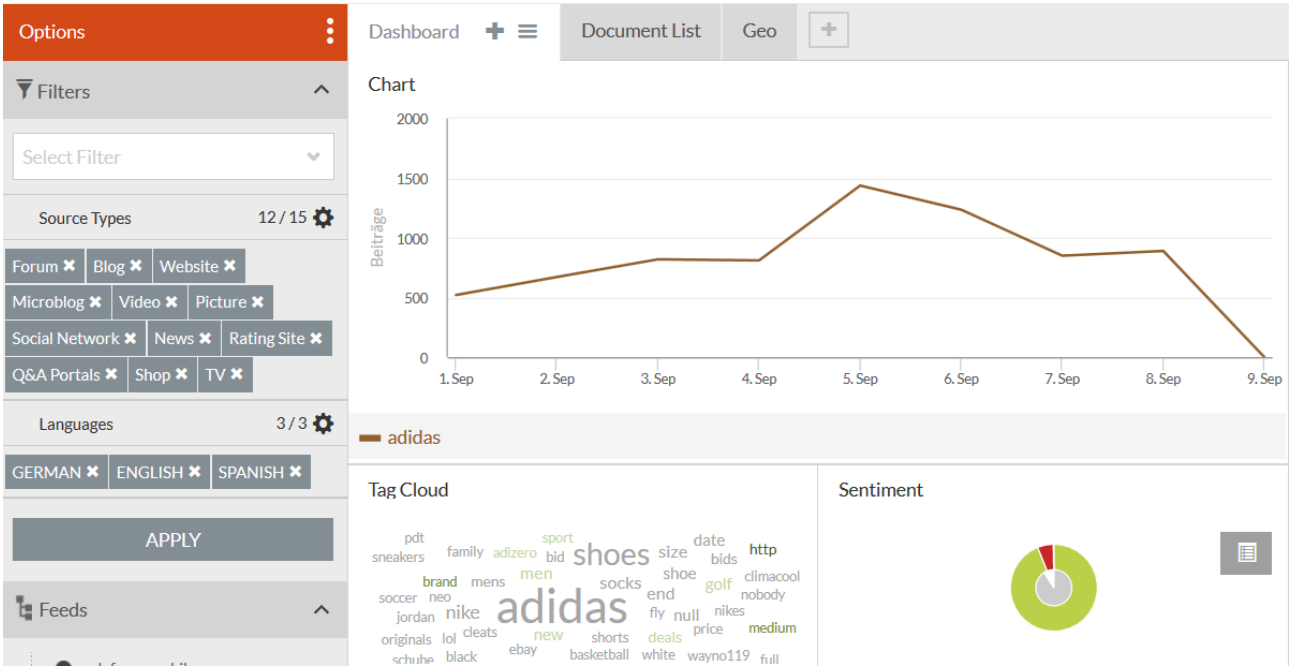

Figure 2: entity tracking "adidas"

4 Conclusion

In this deliverable, we presented the xLiMe approach on text annotation improved for social media as compared to XLike. First, we identified a knowledge base that suits the text annotation in multiple languages. Then with the help of knowledge base, we built a prototype using a sophisticated graph algorithm. The algorithm uses personalized page rank to infer the probability scores for annotations on any unstructured text. The results obtained using the approach is visualized using a sample output.

The present approach is vulnerable to short texts like tweets. Even though the precision of identification of an entity is high, due to noisy nature of short texts recall is low. In the future, we aim to improve the existing techniques to identify the mentions of entities efficiently in the short texts.

Also, we aim to use other techniques like named entity recognition (NER); part of speech tags (POS) etc. to identify the mentions in documents. Evaluation of the approach will be done by creating a sample dataset using crowd sourcing.

5 Acknowledgements

We would like to thank Olga Schamber for her help in improving the presentation of the deliverable.

References

[1] http://www.xlike.org/ (XLike)

[2] http://dbpedia.org/About (DBpedia)

[3] xLiMe Deliverable D1.1 *Meta Data Model*

[4] http://www.vico-research.com/ (VICO)

[5] Taher H. Haveliwala. 2002. Topic-Sensitive PageRank. In Proceedings of the Eleventh International World Wide Web Conference

