Survey Results: Requirements and Use Cases for Linguistic Linked Data

1 Introduction

This survey was conducted by the FP7 Project LIDER (http://www.lider-project.eu/) as input into the W3C Community Group on Linked Data for Language Technology (LD4LT - http://www.w3.org/community/ld4lt/).

This LD4LT Community Group is an open forum where current and potential users of linguistic data can assemble use cases and requirements for Language Technology Applications that use Linked Data. The results will be used to guide future interoperability, research and development activities spanning the language technology and linked data domains.

Potential users are companies and public bodies involved in content management, the language services and localisation industry and other applications of content analytics techniques used in search, recommender systems, sentiment analysis and terminology management.

The main goal of this survey was to understand current industrial needs, requirements and use cases that will help define a roadmap for future R&D activities in multilingual/multimedia content analytics. The survey was made available online via the W3C, recruiting participants by email using our contacts as well as on different mailing lists.

A total of 27 participants were recruited for this survey. The questions considered in our survey are organized in four main parts. The parts are the following: participant profile, NLP application areas, use of language resources, and awareness/maturity in using linked data.

2 Participant Profile

The first part of the survey is concerned with gathering information about the profile of each participant. Participants were asked about the type of organization they are associated with and the industry sector they are active in, allowing them to choose between multiple options. While circulating this survey, we specifically stated our interest in industry participation. Therefore the majority of responders are affiliated with SMEs or large organizations, and only a smaller number with public sector organizations, as can be seen in Table 1. Each participant can have more than one affiliation and can be active in multiple industry sectors. Therefore, participants were allowed to choose more than one option in both cases.

Organization type	Responders
SME	13
Large Organization	6
Public Sector	6
Non-profit	1
Freelancer	0
Other	1

Table 1. Organisation type

The second question about participant profile was related to industry sectors. A list of industry sectors was provided to the participants, but they were also given the option to choose a miscellaneous category, called “Others”. Table 2 presents their responses, showing a marked interest from professionals in Public Sector Publishing, Media, News and Journalism and Localization. Other sectors that showed interest in the area of using linked data in NLP applications for content analytics included the Pharmaceutical sector, Service/Product vendors and eHealth.

Industry sector	Responders
Other	10
Public Sector publishers	9
Media, News and Journalism	8
Localization	7
Pharmaceutical	6
Service / Product vendors (customer support)	6
eHealth	5
Content Management Tool Vendors	3
Libraries, Museums, Digital Humanities	3
Finance	2
ePublishing / eBook	2
eEnergy	1
eTransport	1
Peer production communities	0

Table 2. Industry sector

3 NLP Application Areas

The second part of the survey is concerned with identifying NLP applications in content analytics that are of interest to the industrial community. Several broad areas of applications were identified, including Discovering and Extracting Information, Understanding Opinion, Data Management, and Monitoring and Forecasting. Overall, the use case that achieved the highest consensus with respect to industrial interest is related to the extraction of information from unstructured data, from the broad area of Discovering and Extracting Information.

The following sections discuss industrial interest in each of these areas separately.

3.1 *Discovering and Extracting Information*

The first application area brings together several use cases related to information discovery and information extraction, which are listed in Table 3. The majority of the responders identified the area of information extraction from unstructured data as an area they are interested in. Other areas that gathered a majority of votes included entity and event detection, expert finding and semantic search.

Use case	**Responders**
Extraction of information from unstructured data	24
Entity and event detection	18
Expert finding from unstructured and structured data	18
Semantic search	18
Text-to-semantics conversion	11
Question answering in natural language	10
Multimedia and video search, visual search	9
Fact validation using unstructured / web data	8
Speech-to-semantics conversion	5

Table 3. Use cases related to discovering and extracting Information

3.2 *Understanding Opinion*

The second application area groups together use cases related to the broad area of understanding opinions. An overview of the responses in this area is presented in Table 4. A large number of participants identified impact analysis as a relevant use case, immediately followed by use cases in sentiment and opinion mining. Other use cases of interest for the industrial community include mining customer interaction data and trend mining, with almost half of the participants expressing interest in them.

Use case	**Responders**
Impact analysis (e.g. of marketing campaigns or other marketing measures)	16
Sentiment / opinion mining	15
Mining customer interaction data to acquire insights about their behavior	13
Trend mining	13
Identifying key opinion holders / opinion leaders	12
Identifying and making explicit the argument structure and logical relation between opinions within public discourse about a topic	11
Identifying (potentially) opposing communities	5
Identifying irony / sarcasm in web texts / reviews	5

Table 4. Use cases related to understanding opinion

3.3 *Data Management*

The area of data management organizes several NLP application areas related to creating, organizing, sharing and storing content and data, which are listed in Table 5.

The largest number of participants showed an interest in use cases related to data integration and content summarization. More than half of the participants expressed an interest in tools that support ontology building, evolution, and maintenance and topic detection.

Use case	Responders
Data integration	17
Content (text, multimedia) summarization	15
Support for text-based ontology building / evolution / maintenance	14
Topic detection	14
Aspect oriented data summarization	13
Rapid knowledge base formation from textual data for analytics task	13
Supporting development of (multilingual) terminologies / thesauri / term bases	13
Taxonomy maintenance	11
Machine translation	8
Speech-to-text conversion	8
Natural language generation from templates, database content etc.	7
Digital preservation of multilingual, multimedia content	6
Information kiosk	6
Multimedia eLearning	6
Speech processing	4
Computer and video games	1

Table 5. Use cases related to understanding opinion

3.4 *Monitoring and Forecasting*

The last broad area considered in this survey is related to monitoring and forecasting topics and entities of interest, described in more detail in Table 6. The most relevant use case to the industrial community was the use case related to predictive analytics over text data, followed by use cases that address tracking entities on the Web.

Use case	Responders
Predictive analytics over text data	18
Tracking entities (people, products) on the Web	15
What-if-simulation based on content analytics results finding relevant communities/fora/discussion pages on the Web	6

Table 6. Use cases related to understanding opinion

4 Use of Language Resources

This part of the survey was concerned with mapping industrial use of existing language resources. Participants were asked about the type of language resource that they make use of in their daily activities, as can be seen in Table 7. Dictionaries, corpora, and tokenizers are the most widely used resources by the industrial community.

Language resource	Responders
Dictionaries (Monolingual / Bilingual / Multilingual)	15
Corpora (Written / Spoken / Multimodal)	13
Tokenizers	12
NLP Frameworks: UIMA / GATE / NLTK Toolkit	11
Part-of-speech Taggers	11
Sentence Splitters	11
Terminologies	11
Encyclopedic resources (DBpedia, YAGO, BabelNet, etc.)	9
Parsers	9
Term bases	9
Translation memories/parallel text	9
Machine Translation Systems (e.g. Moses, Google, Bing, …)	8
Others	5

Table 7. Type of language resource

The second question related to the use of language resources was concerned with the location of the language resource used. The majority of the participants make use of a mixture of language resources that are produced within their organization together with external language resources, as can be seen in Table 8.

Language resource location	Responders
In-house	6
External language resources	4
Both of the above	17

Table 8. Location of language resource

5 Awareness/maturity in using Linked Data

The last part of the survey gathers information about the awareness and maturity of using Linked Data and Linguistic Linked Data, in Tables 8 and 9, respectively. Not surprisingly, the majority of the participants are very aware of Linked Data.

Awareness	Responders
Very aware	16
Not so	8
Not at all	3

Table 9. Linked Data usage

The same thing cannot be said about Linguistic Linked Data, because less than half of the participants stated that they are aware of this resource.

Awareness	Responders
Very aware	12

Not so	9
Not at all	6

Table 10. Linguistic Linked Data usage