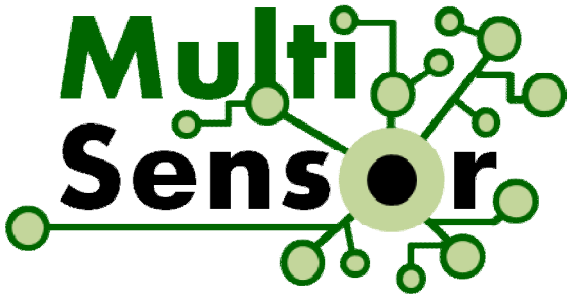


MULTISENSOR

Mining and Understanding of multilinguaL contenT for Intelligent Sentiment Enriched coNtext and Social Oriented inteRpretation

FP7-610411

D6.1

Summarisation infrastructure and baselines

Dissemination level:	Public
Contractual date of delivery:	Month 11, 30/09/2014
Actual date of delivery:	Month 11, 30/09/2014
Workpackage:	WP6 Summarisation and content delivery
Task:	T6.1 Basic summarisation infrastructure T6.2 MULTISENSOR summarisation dataset
Type:	Report
Approval Status:	Final
Version:	1.1
Number of pages:	23
Filename:	D6.1_SummarisationInfrastructure_2014-09-29 v1.1.pdf

Abstract

This deliverable reports the implementation of an extractive summarisation baseline for both single and multiple-document summarisation, as part of task T6.1, and the progress towards the creation of a summarisation dataset, as part of task T6.2.The dataset comprises a paired corpus of press articles and hand-crafted summaries, a corpus of press articles annotated automatically with linguistic information and links to semantic content, and a corpus annotated manually with linguistic information. These three corpora are being used

to develop different summarisation techniques and are currently in different stages of advancement.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	15/07/2014	Layout proposal	G. Casamayor (UPF)
0.2	8/07/2014	First draft	G. Casamayor (UPF)
0.3	9/07/2014	Comments	V. Aleksić (LT)
0.4	15/07/2014	Changes in response to feedback from Vera Aleksić.	G. Casamayor (UPF)
0.5	1/09/2014	Comments and suggestions	A. Moumtzidou, S. Vrochidis (CERTH)
0.6	15/09/2014	Changes in response to feedback from Stefanos Vrochidis and Anastasia Moumtzidou.	G. Casamayor (UPF)
0.7	26/09/2014	Version for internal review. Added description of licensing issues for the publication of the summarisation dataset.	G. Casamayor (UPF)
0.8	28/09/2014	Internal review	V. Aleksić (LT)
1.0	29/09/2014	Final version after feedback from internal review by Vera Aleksić.	G. Casamayor (UPF)
1.1	29/09/2014	Final updated version after minor corrections.	G. Casamayor (UPF)

Author list

Organisation	Name	Contact Information
UPF	Gerard Casamayor	gerard.casamayor@upf.edu
UPF	Simon Mille	simon.mille@upf.edu
DW	Tilman Wagner	tilman.wagner@dw.de
PR	Mirja Eckhoff	mirja.eckhoff@pressrelations.de
PIMEC	Teresa Forrellat	TForrellat@pimec.org

Executive Summary

This report presents a basic approach for text summarisation which constitutes a baseline for future efforts in text summarisation. This approach involves applying statistical methods applied to an aligned corpus of texts and hand-crafted summaries in order to obtain metrics for the assessment of relevance of text sentences. The generated summaries are the result of evaluating the original texts and composing a text by extracting verbatim the most relevant sentences and joining them together. A state-of-the-art extractive summarisation toolkit has been used to implement the baseline which supports both the summarisation of a single document of a set of documents. The implementation is being integrated into the general architecture of the MULTISENSOR project and is also being tuned to each specific scenario separately.

A novel summarisation approach will be developed that is based on abstractive summarisation techniques. Instead of copying fragments from the original texts with no understanding of their meaning, the texts will be analysed using Information Extraction (IE) and Natural Language Processing (NLP) methods, as part of WP2. The extracted contents will be stored in a semantic repository, integrated with information from audiovisual content, linked to other datasets, and enriched through reasoning (WP2, WP4 and WP5). Starting from the contents in the semantic repository, WP6 will address the creation of abstractive summaries using Natural Language Generation (NLG) methods. A first step for the development of NLG methods is the compilation of a summarisation dataset that can be used to obtain heuristics or train the various tasks involved in the generation of a summary from data.

The current progress in the compilation of this dataset is reported in this document. The dataset comprises three separate corpora: (i) a corpus for the extractive summarisation approach which consists of pairs of texts and human-authored summaries, (ii) a corpus of texts automatically annotated with linguistic information and the contents extracted from them, and (iii) a corpus of texts manually annotated with linguistic features. The second and third corpora are used to develop specific tasks of the NLG pipeline in charge of generating abstractive summaries. Corpus (ii) is used for the extraction of heuristics for the production of a text plan from the contents in the semantic repository. It is also used to automatically derive dictionaries to render the text plan in (multilingual) natural language. The task of producing natural language from the text plan is further leveraged with the corpus (iii) annotated with syntactic and morphologic structures.

Abbreviations and Acronyms

ANNIE	A Nearly-New IE system
API	Application Programming Interface
AS	Automatic Summarisation
CoNLL	Conference on Natural Language Learning
GATE	General Architecture for Text Engineering
IDF	Inverse Document Frequency
IE	Information Extraction
JAPE	Java Annotation Patterns Engine
JSON	JavaScript Object Notation
JSON-LD	JSON for Linking Data
LOD	Linked Open Data
MT	Machine Translation
MTT	Meaning Text Theory
NE	Named Entity
NER	Named Entities Recognition
NIF	NLP Interchange Format
NLG	Natural Language Generation
NLP	Natural Language Processing
OWL	Web Ontology Language
RDF	Resource Description Format
REST	Representational State Transfer
UC1	Use Case 1
UC2	Use Case 2
WP	Work Package

Table of Contents

1 INTRODUCTION

This deliverable describes the work done in Work Package 6 (WP6) during the first 11 months of the project MULTISENSOR. As the title indicates, this deliverable covers the set up of a web-based infrastructure that supports an initial implementation of an automatic summarisation (AS) system based on state-of-the-art summarisation tools. These tools will serve as the baseline for the more advanced AS technologies developed in the scope of WP6. The summarisation infrastructure described here is integrated in the project general architecture and contributes to milestone 2 (MS2) of the MULTISENSOR project, which foresees the completion of a first operational system. The work described in this document contributes towards the completion of tasks 6.1 (Basic summarisation infrastructure) and 6.2 (MULTISENSOR summarisation dataset). These tasks correspond to the activity A.6.1 Extractive summarisation, described in the project roadmap D7.1 and scheduled to be completed by the end of the first year, one month after the release of this deliverable (see Figure 1).

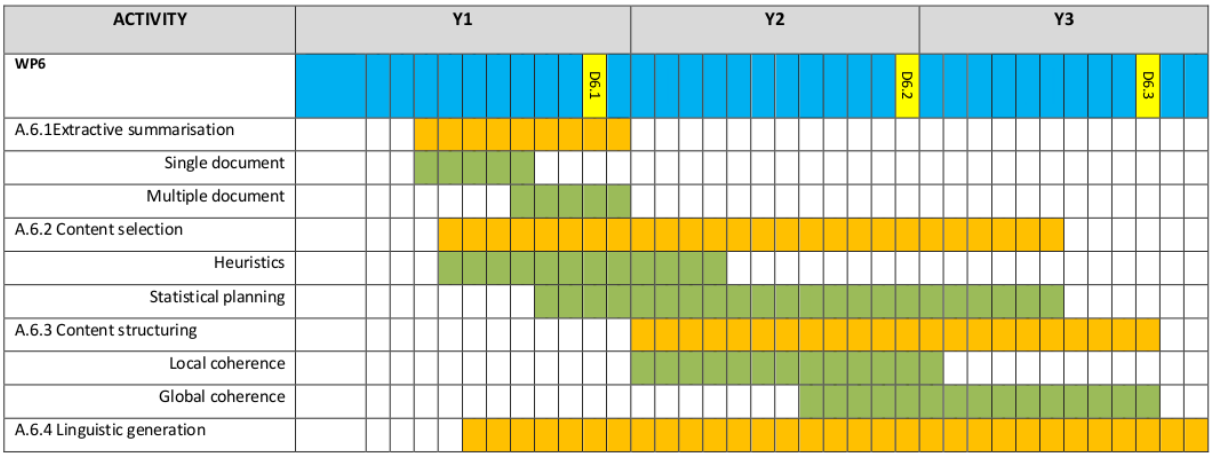


Figure 1: Timeline of activities, reproduced from D7.1

The basic summarisation infrastructure consists of a multilingual text-to-text extractive summarisation module. Extractive summarisation addresses the production of summaries from textual sources by identifying relevant fragments (e.g. paragraphs, sentences, phrases) in the sources and using them to compose a summary. Extractive summarisation is seen as opposed to abstractive summarisation where information is extracted from the sources and used to generate a new text from scratch using NLG methods. The advanced AS methods researched in tasks 6.3, 6.4, 6.5 and 6.6 follow the abstractive approach and will be eventually compared against the extractive summarisation strategy described in this document.

State-of-the-art methods for extractive summarisation (Gupta and Lehal 2010;Nenkova and McKeown 2012) exploit paired corpora of documents and human-authored summaries by applying statistical and machine learning methods. The summaries constitute examples of what contents are judged most relevant in the original documents, and the success of the AS approaches largely depends on their quality and number of pairs of document-summary available. Task 6.2 is about compiling this aligned corpus plus any other textual material that could be useful for latter stages of the work package.

This document is divided into a section describing the AS infrastructure (Section 2), the summarisation corpus or dataset (Section 3) and a closing section, where we draw some conclusions about the work done so far (Section 4).

2 BASIC SUMMARISATION INFRASTRUCTURE

2.1 Determining the user requirements

In order to determine the requirements of the summarisation module, we conducted a simulation exercise together with the user partners (pressrelations, Deutsche Welle, PIMEC). The goal was to determine user expectations and requirements on the summaries generated by the MULTISENSOR system. We came up with a list of types of summaries and information presentation strategies together with mock-ups created ad-hoc for each use case and scenario:

List of keywords: this type of summary is a ranked list of multi-word expressions taken literally from the text on the basis of their relevance. Our lists were determined automatically using a statistical measure of relevance based on their frequency. Lists can be presented to end users as a plain list, as a cloud, etc.

Extractive summary: This type of summary is a text composed of fragments extracted verbatim from the original document. The extractive summaries were obtained automatically using the SUMMA toolkit[1] and were based on the extraction of phrases and of full sentences.

Abstractive summary: An abstractive summary is a text produced using NLG methods from data extracted from other texts. Abstractive summaries have improved readability and coherence over extractive summaries. Additional information not found in the source document(s) can be added to the summary, such as the results of sentiment and opinion analysis, data obtained through inference or data from other open linked datasets. Our simulated abstractive summaries were hand-crafted. Abstractive summaries may be presented to end users as plain text or enriched with links and semantic annotations of the data used to generate them.

Graphic visualisation of the semantic repository: Instead of an abstractive summary, the very contents from which the summary is generated can be displayed directly to the user using some graphical interface. In the MULTISENSOR system this would be equivalent to exposing selected fragments of the semantic repository. Two visualisation mechanisms were described: customised html layouts as in Google's Freebase[2]and navigable/interactive graphs.

A document describing these alternatives and the mock-ups were presented by the user partners to potential end users of the MULTISENSOR system. The results of these interviews together with the insights of the partners themselves were discussed together with the technical partners in the third plenary meeting which took place in Barcelona (M7). The user interviews are described in detail in D8.2.

[1]See http://www.taln.upf.edu/pages/summa.upf/index.htm

[2] http://www.freebase.com/

2.2 Implementation of the summarisation service

The extractive summarisation component is implemented with SUMMA, a text summarisation toolkit based on the GATE[3] framework for NLP. SUMMA provides resources and tools to perform statistical analysis of texts and build multidocument and multilingual extractive text summarisation applications based on the result of such analysis. Following the service-oriented architecture of the MULTISENSOR system (see D7.1), the extractive summarisation system is deployed as a REST web service with a public API to exchange JSON-based messages. The AS service is executed towards the end of the main text analysis pipeline of the MULTISENSOR system, as depicted in Figure 2:

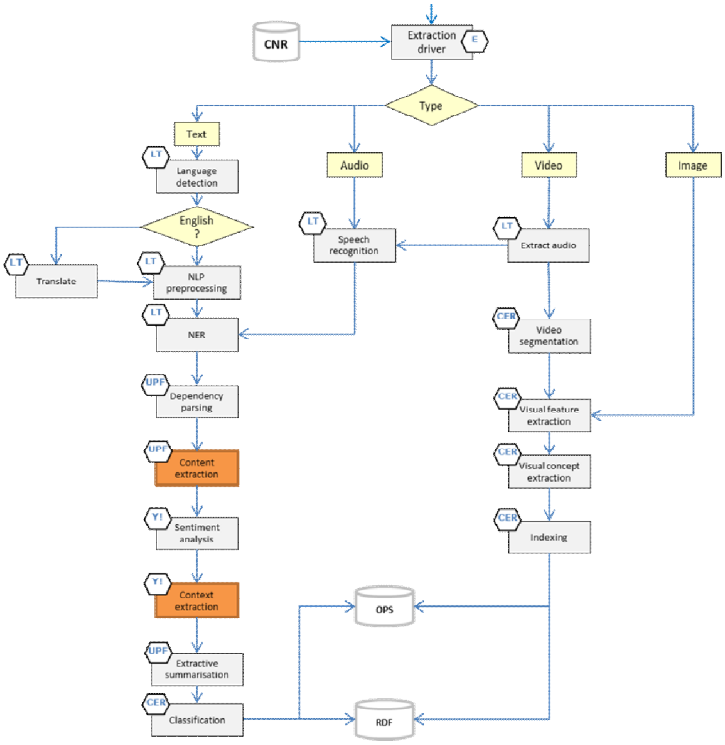


Figure 2: Text analysis and summarisation pipeline

Two SUMMA-based pipelines have been created to handle single-document and multi-document summarisation respectively. The single-document pipeline executes the following SUMMA and ANNIE[4] modules in sequence:

1. ANNIE Tokenizer: determines documents token and annotates them.
2. JAPE Transducer: processes punctuation marks and splits words containing apostrophes and other marks.
3. ANNIE Sentence splitter: determines sentence boundaries and annotates them.
4. SUMMA NEs statistics: calculates and annotates basic statistics for each token:
    a. Number of times it appears in the document, sentence and paragraph.

[3] See https://gate.ac.uk/

[4] An information extraction toolkit bundled with GATE, see https://gate.ac.uk/sale/tao/splitch6.html#x9-1280006

   b. Inverted document frequency (idf), as measure of how much information a term provides, based on how common it is in a collection documents.
   c. Term frequency multiplied by inverted document frequency (tf*idf).
5. SUMMA Term Frequency Filtering: filters out tokens which have a tf*idf below 10.
6. SUMMA Vector Computation: creates a feature vector for the whole document containing pairs of tokens and tf*idf values.
7. SUMMA Normalise Vector: normalises the text vector to [0..1].
8. SUMMA Vector Computation: creates a feature vector for each sentence containing pairs of tokens and tf*idf values.
9. SUMMA Normalise Vector: normalises the sentences vectors to [0..1].
10. SUMMA Position scorer: adds a feature to each sentence indicating its relevance according to its position in the text.
11. SUMMA Sentence document similarity: calculates and annotates for each sentence its similarity to the whole document according to their respective tf*idf vectors.
12. SUMMA Sentence term frequency scorer: sums and annotates the global frequencies in the text of each term in the sentence.
13. SUMMA First sentence similarity: calculates and annotates for each sentence its similarity to the first sentence according to their respective tf*idf vectors.
14. SUMMA Simple summariser: performs a weighted combination of all sentence metrics calculated so far in order to obtain a relevance score for each sentence. Then a subset of the most relevant sentences can be extracted to generate the summary. The following metrics are used:
    a. Similarity to first sentence
    b. Similarity to whole document
    c. Sentence term frequency
    d. Position score

The architecture for the summariser of multiple documents analyses and calculates metrics for each sentence of each document in the same way as the single document pipeline. It then performs additional calculations to compare each sentence to the whole set of input documents, as described below:

1. ANNIE Tokenizer
2. JAPE Transducer
3. ANNIE Sentence splitter
4. SUMMA NEs statistics
5. SUMMA Term Frequency Filtering
6. SUMMA Position scorer
7. SUMMA Vector Computation for whole document
8. SUMMA Normalise Vector for whole document
9. SUMMA Vector Computation for each sentence
10. SUMMA Normalise Vector for each sentence
11. SUMMA Sentence document similarity
12. SUMMA Sentence term frequency scorer
13. SUMMA First sentence similarity
14. SUMMA Centroid Computation: calculates the central point in the feature space using the vectors of each document.

15. SUMMA Centroid Sentence Similarity: each sentence is annotated with its similarity to the document centroid calculated in the previous step.
16. SUMMA Simple Summariser: the weighted combination of features for each sentence uses the same metrics as in the single document pipeline with the addition for the similarity to the centroid.

Both pipelines can be parameterised with the following:

- A compression rate based either on number of tokens or number of sentences.
- A set of weights for the metrics used in the simple summariser module.
- An IDF table obtained from a corpus of relevant documents.

The pipelines have been tuned to the journalistic genre with an IDF table automatically obtained from a generic corpus of 65.000 articles provides by pressrelations. At the moment all metrics are set with identical weights in both pipelines. These summarisation pipelines implement basic AS strategies which can be extended to match the performance of state-of-the-art extractive summarisation systems. The following steps are required to achieve real state-of-the-art performance:

- The tokenisation and sentence splitting modules must be replaced with the annotations generated by Linguatec services.
- Separate pipelines must be set up which are specifically tuned to each use case and scenario.
- Additional modules are being considered for each pipeline that exploit the idiosyncrasies of the documents belonging to each use case, and also the information made available by the text analysis services in WP2. The modules being experimented with include, amongst others, sentence-to-title and paragraph-to-title similarity metrics, semantic analysis based on named entities, and metrics based on n-grams.
- Human-authored summaries provided by other partners should be used to empirically set the weights for the various metrics. This will to be done following the procedure detailed in Brügmann et al. (2014).

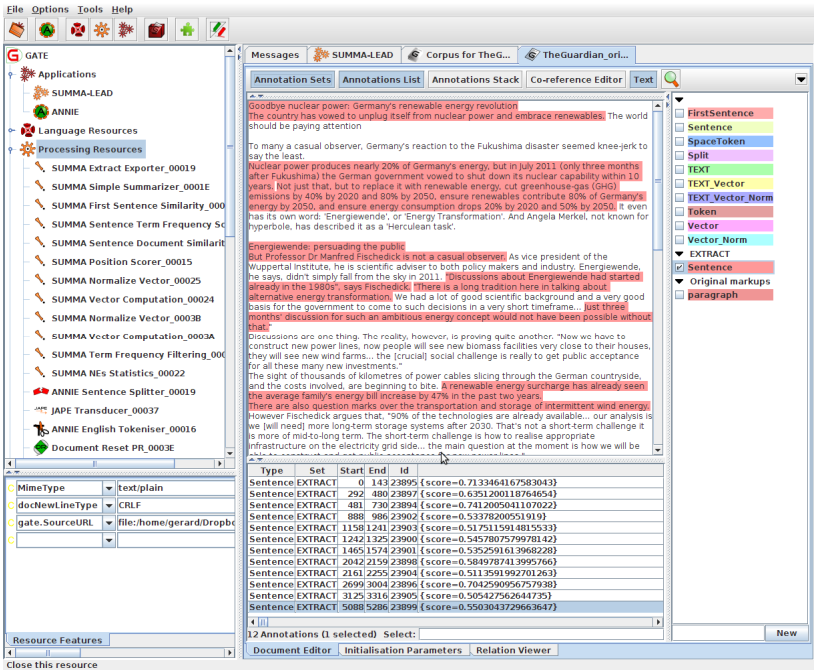


Figure 3: Single-document summarisation pipeline in GATE

3 MULTISENSOR SUMMARISATION DATASET

State-of-the-art and advanced methods for the automatic production of summaries from texts or data are mostly based on empirical evidence extracted automatically from corpora of texts. The following subsections describe the corpora compiled as part of task T6.1 and their use in the development of AS tools.

3.1 Extractive summarisation corpora

Methods for extractive summarisation require a corpus of texts representative of type of documents to be summarised paired with human-crafted summaries. This paired corpus is analysed using NLP methods and metrics are extracted that determine what fragments of any new texts should be included in the summary. Most of the techniques used to analyse the text are shallow and therefore sensitive to the language in which texts are written. For this reason, when dealing with multilingual input texts, separate versions of the training corpus are needed containing source documents and their summaries in each language.

Given the difficulties faced in obtaining a significant amount of human-authored summaries for each language and scenario, it was decided that, as a workaround, the MT technologies developed by Linguatec in the scope of the MULTISENSOR project would be used to translate all input texts into English, and the summaries back to the original language of the document or the language requested by the user. This way partners are be able to focus from now on the production of summaries tailored to each use case. While MT translation from all languages to English has already been released, translation of the English summaries to other languages will be available when Linguatec releases support for additional languages during the second year of the project.

As a result of the errors introduced by the MT in the translated summaries, the performance of the baseline extractive summarisation system will suffer when applied to translated texts. Our extractive methods are particularly sensible to incorrect or inconsistent translation of frequent terms. Depending on the quality of the translations for each combination of scenario and language, it may be necessary to explore mechanism to ameliorate the situation in some of these combinations. A possible mechanism is to modify the summarisation pipeline so that it relies less on language-specific terms and more on the language-independent results of the text analysis services developed in the context of WP2, namely named entities and other contents marked in the texts. This and other mechanisms will be considered in future time.

A first corpus of general press articles was made available by pressrelations and stored in the Elastic Search[5] news repository maintained by Everis. It contains 65.642 pairs of news articles and human-crafted summaries. The summaries have an average length of 125 words against an average length of 445 words in the original documents. This corpus has been used to tune the basic single and multiple document summarisers described above.

[5] http://www.elasticsearch.org/

UPF determined the exact requirements for the corpora needed to further tune the AS tools for each scenario:

- 250 pairs of articles and summaries in English for each scenario, that is, 750 pairs in total.
- The summaries must have an approximate length of 200 words.
- The writers can replicate parts of the original document or use their own words to write the summaries.
- The articles must be fairly recent (up to 2 years old, approx.) in order to guarantee that they cover recent actors and terminology in each use case.

pressrelations already contributed a set of 1.900 pairs of texts and summaries belonging to the UC1 scenario 2 (household appliances). The summaries have an average length of 107 words against an average article length of 527 words. This dataset is being used by UPF to train and deploy a scenario-specific AS pipeline.

Deutsche Welle and PIMEC are investing a person month each (T6.2) in the creation of summarisation sets for their respective scenarios, UC1 scenario 1 (energy policies) and UC2 scenario 1 (internationalisation). As soon as these two datasets are available, UPF will use them to tune and deploy new scenario-specific AS pipelines.

3.2 Abstractive summarisation corpora

Two different datasets are used for the abstractive summarisation approach. A first automatically annotated corpus containing both linguistic and text-to-data annotations is used to obtain metrics for the planning of texts from the contents of the MULTISENSOR semantic repository, and for the automatic derivation of dictionaries used to render the contents as multilingual text. A second dataset consists of manually annotated multilingual texts with syntactic and morphologic information and its purpose is to train the statistical methods to map text plans to multilingual text.

3.2.1 Text planning and lexicalisation dataset

Abstractive summarisation applies NLG methods to generate multilingual textual summaries (accompanied by multimedia if needed)from data. In MULTISENSOR, the starting point of the abstractive summarisation is the contents in the semantic repository, which are obtained from text documents using NLP and IE methods as part of WP2, and then integrated with data from other sources (images from multimedia documents, videos, social media, etc.) and enriched through reasoning as part of WP4 and WP5. With the purpose of reusing in WP6 the linguistic information and semantic data extracted from the texts in WP2, it has been agreed that all text analysis components produce annotated text as their output, and that these annotations are kept together with the analysed text as an annotated corpus. So far, the following annotations are foreseen:

1. Layout annotations (e.g. title, lead) detected by the generic crawler by Yahoo (WP7) using Boilerpipe[6] or similar.

[6] See https://code.google.com/p/boilerpipe/

2. Tokens, sentences, named entity types and disambiguated links to entities in the datasets in the semantic repository, returned by the NER service maintained by Linguatec (WP2).
3. Surface- and deep-syntactic parses of each sentence, returned by the dependency parser service maintained by UPF (WP2).
4. Coreference links for nominal expressions, returned by the coreference resolution service maintained by UPF (WP2).
5. Links from nominal expressions to domain-specific entities and concepts and relations between entities indicated by verbal predicates, both returned by the content extraction service maintained by UPF (WP2).
6. References from nominal expressions and predicates to their corresponding entries in general online lexical resources and terminological glossaries. At the moment the resources being considered are BableNet[7], WordNet[8], EuroVoc[9], Reegle Glossary[10], VerbNet[11], PropBank[12] and FrameNet[13]. UPF is in charge of this annotation, which is described in D2.2 (WP2).
7. Annotations of sentiments associated to mentions of entities as returned by the sentiment analysis module maintained by Yahoo (WP3).
8. Contextual information associated to the document as returned by the context extraction module (WP3).
9. Concepts and events depicted by images and/or video keyframes extracted in a supervised manner (WP2).

It has also been agreed that a common representation for the linguistic annotations based on the NLP Interchange Format (NIF)[14]will be used. NIF is a stand-off linguistic annotation format modelled using a set of OWL ontologies. It facilitates the creation and publication of RDF-based annotated corpora in the LinguisticLOD Cloud[15]. The NIF-based messages exchanged by the linguistic analysis services are serialised using JSON-LD[16]. The resulting annotations and texts constitute a NIF open LOD corpus containing annotations of diverse nature.

A typical NLG system is implemented using a pipeline architecture comprising the following modules:

- Content determination: selection of the input data to be communicated in the text.

[7] See http://babelnet.org/

[8] See http://wordnet.princeton.edu/

[9] See http://eurovoc.europa.eu/drupal/

[10] See http://www.reegle.info/glossary

[11] See http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[12] See http://verbs.colorado.edu/~mpalmer/projects/ace.html

[13] See https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=home

[14]See http://persistence.uni-leipzig.org/nlp2rdf/

[15] See http://linguistics.okfn.org/resources/llod/

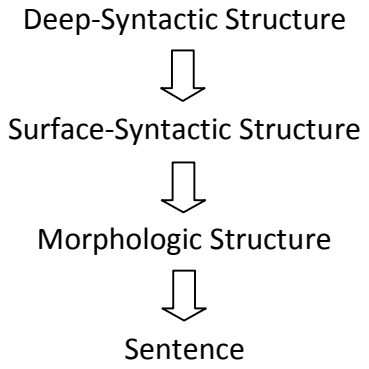
[16] See http://json-ld.org/

- Discourse structuring: organisation of the selected contents into a coherent whole, including the determination of the order in which contents are communicated.
- Sentence planning: also known as microplanning; involves determining the structure of the sentences to be generated and includes subtasks like referring expression generation, content aggregation into sentences and lexicalisation.
- Surface realisation: the production of actual text from the planned text and sentences, which consists mainly in introducing all idiosyncratic elements, ordering the words, and managing the morphological interactions between them (agreements, compositions, contractions, etc.).

There are plans for the extraction of metrics from the NIF-annotated corpus for content determination (T6.3) and lexicalisation (T6.4 and T6.5), based on the annotations of contents (links to entities, relations, sentiments and context) and references to lexical resources respectively. At the moment, preliminary work is being conducted to assess the viability of automatically deriving heuristics for the content determination task and lexical dictionaries for the lexicalisation subtask. D2.2, due in month 12, will describe an experimental annotation of texts with freely available online lexical resources and data which should contribute towards the implementation of the WP2 pipeline but also towards tasks T6.3, T6.4 and T6.5.

3.2.2 **Multilingual surface generation dataset**

For surface generation (i.e., sentence planning and surface realisation), two different approaches are foreseen, both based on a pipeline of graph transducers which convert the output of the text planning stage into a well-formed text, and both being implemented as part of the Mate tools[17]. The generation is performed step by step, following the layers of the Meaning-Text Theory (Mel'čuk, 1988):



The first approach consists in manually crafting graph-transduction grammars for each transition between two consecutive layers. In combination with the rules, dictionaries of three different types are required: one that includes equivalences between the entities in the semantic repository and the words of the different languages of the project (semantic

[17] http://code.google.com/p/mate-tools/

dictionary), one that describes syntactic properties of these words (lexical dictionary), and one that contains the inflection patterns of each word (morphological dictionary). As mentioned before, UPF will attempt the automatic derivation of the dictionaries from the annotated data and references to existing open lexical resources in the Text Planning and Lexicalisation Dataset. A new graph transduction platform containing advanced rule and dictionary editors is being implemented and tested by UPF as part of the Mate tools.

The second approach consists in performing each transition with statistical modules trained on annotated data. Indeed, by aligning node by node a parallel corpus of two consecutive levels of representation, it is possible to apply Machine Learning techniques and obtain models for a statistical generator. **For this purpose, a multilingual corpus containing MTT-based linguistic annotations is being manually annotated by UPF. We refer to this dataset as the multilingual surface generation dataset.**

For each language, we need morphologic, surface-syntactic and deep-syntactic annotations, as illustrated by Figure 4, Figure 5 and Figure 6 respectively. All the layers (and the original text) need to be aligned sentence by sentence and node by node, which can be done thanks to unique identifiers associated to each sentence and node. The morphologic annotation consists of features associated to each word of the sentence, including coarse-grained and fine-grained part-of-speech, gender, number, tense, aspect, finiteness, mood, person, etc. On this layer only, the order of the words is kept. The surface-syntactic annotation consists of dependency trees with all the words of a sentence linked by idiosyncratic relations. At the deep-syntactic layer, all functional (i.e. non-meaningful) units are removed: definite and indefinite determiners and auxiliaries are replaced by attribute/value pairs on the concerned nodes, while punctuations and governed prepositions and conjunctions are simply removed. In addition to attributive and coordinative relations, the dependency relations also encode predicate-argument information, through the assignment of an argument slot in the valency (*subcategorisation framework*) of its governor predicate (see *I* and *II* in Figure 6).

The producer *thinks* *that the new song will be successful soon.*
[…] […] *[lemma=think* *[…][…] […] […] […] […][…] […][…]*
PoS=verb
finiteness=finite
tense=present
mood=indicative
person=3
number=singular]

Figure 4: Morphologic structure

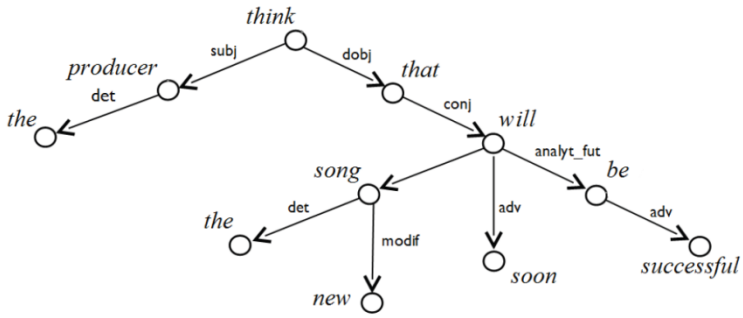


Figure 5: Surface-syntactic structure

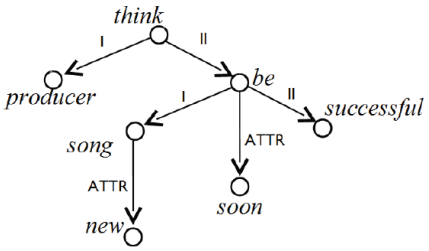


Figure 6: Deep-syntactic structure

We currently have available multilayered corpora for Spanish and English, at different stages of advancement; German is being tackled and French and Bulgarian will be addressed in the next weeks.

In Spanish, we use the Angora-UPF dependency Treebank (Mille et al., 2013) in CoNLL format[18]. In the Ancora-UPF, each layer is already independently annotated, and the annotation has been manually validated, with a high inter-annotator agreement. For the surface-syntactic annotation, there are several dependency tag sets available, with many or few tags. The different dependency tag-granularities will be tested in order to find the optimal tag set for the statistical generation. The treebank contains 3.513 sentences (100.892 tokens at the surface-syntactic and morphologic layers, 65.889 tokens at the deep-syntactic layer).

[18]The corpus underlying both treebanks is the same: Ancora from 2008 (Taulé et al., 2008).

In English, we use the Penn Treebank 3 (Marcus et al., 1994) converted to dependency trees (Johansson and Nugues, 2007), which we use as such as surface-syntactic and morphologic annotations. We automatically derived a first version of the deep annotation from the surface annotation using graph transduction grammars implemented in the Mate environment. During the mapping, we removed all determiners, auxiliaries, *that* complementisers, infinitive markers *to*, punctuations and a reduced list of governed prepositions. In order to obtain the list of prepositions to remove, we ranked all predicates of the corpus based on frequency, and for the predicates that appear at least 150 times, we checked manually if they have some governed elements in the corresponding frameset of the Unified Verb Index.[19]We found 152 different predicates which govern one or more prepositions. The predicates have from one to four slots which can require a preposition, and from one to four different prepositions per slot. Then, we built up the list of prepositions to be removed based on the governing predicate, the argument slot and the name of the preposition. The treebank currently contains41.678 sentences (1.015.843 tokens at the surface-syntactic and morphologic layers, 768.865 tokens at the deep-syntactic layer). The annotation is currently being reviewed and will be improved.

In German, we follow the same method as for English; we start with the TIGER corpus (Brants et al., 2002), from which we automatically derive the deep-syntactic annotation thanks to graph transduction grammars. Since the surface-syntactic annotation is quite different from the English one–the dependency tag set and the annotation scheme are not the same – it is not possible to re-use the English mapping. The German grammars are currently incomplete.

[19]http://verbs.colorado.edu/verb-index/index/L.php, or http://verbs.colorado.edu/propbank/framesets-english/

4 CONCLUSIONS

This document describes the progress of WP6 in the first 11 months of the MULTISENSOR project. The main tasks concerned are the implementation and set up of a basic summarisation architecture and the compilation of a dataset for the summarisation task. An initial architecture based on extractive AS methods is ready for both single and multiple document summarisation of English texts. In the next couple of months the implementation should be further adjusted to meet a state-of-the-art level. The summarisation of texts in other languages than English, however, due to the absence of annotated corpora, will be delayed until the start of Y3 when the MT technologies being developed by Linguatec become available, which will be used as a workaround. Another factor which may cause potential delay is the availability of scenario-specific corpora, which so far has been made available only for UC1-scenario 2 (household appliances) and, to a lesser degree, to UC1-scenario 1 (energy policies).

The summarisation dataset for abstractive summarisation comprises a Text Planning and Lexicalisation Datasetand a Multilingual Surface Generation Dataset. The first is a NIF-encoded LOD corpus that is produced automatically by the WP2 and WP3multimedia analysis pipeline. This corpus, when available, will be used to help develop methods for the content selection and lexicalisation tasks of the NLG pipeline in charge of generating abstractive summaries. The other datasets a manually annotated MTT-based multilingual corpus used to train the MATE statistical generator for the surface generation task of the NLG pipeline.

The publication of the summarisation datasets is subject to licensing restrictions for some corpora. Thus, the extractive corpora are to be published on-line using URLs pointing to the original texts to avoid legal issues. The Text Planning and Lexicalisation Dataset will be released when the text analysis pipeline in development as part of WP2 is ready and only for those texts which are open domain. In the case of the Multilingual Surface Generation dataset, we do not own any of the rights for the third-party multilingual corpora annotated with surface linguistic data. Consequently, it will not be possible to release this dataset. WP6 partners are considering as a possible workaround the release of the tools (as, e.g., Johansson and Nugues, 2007) which convert the surface corpora into the deep corpora, so that anyone in possession of a license for a corpus can produce the deep corpora easily.

5 REFERENCES

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. 2002. “The TIGER treebank. In Proceedings of the workshop on treebanks and linguistic theories”, vol. 168.

Gupta, V., &Lehal, G. S. 2010.“A survey of text summarisation extractive techniques”, Journal of Emerging Technologies in Web Intelligence, vol. 2(3), pp. 258-268.

Johansson, R., &Nugues, P. 2007. “Extended constituent-to-dependency conversion for English”, In Proceedings of Nodalida 2007, pp. 105-112.

Mel'čuk, I. 1988.“Dependency syntax: Theory and practice”, State University of New York Press.

Mille, S., Burga, A., &Wanner, L.. 2013. “AnCora-UPF: A Multi-Level Annotation of Spanish”,In:Proceedings of the Second International Conference on Dependency Linguistics (DepLing‘13), pp. 217-226. Prague, Czech Republic.

Nenkova, A., &McKeown, K. 2012. “A survey of text summarisation techniques”, In Mining Text Data, pp. 43-76.Springer US.

Saggion, H. 2008. “A Robust and Adaptable Summarisation Tool”, TraitementAutomatique Des Langues, vol. 49(2), pp. 103–125.

Taulé, M., Martí, M. Antònia, & Recasens, M.2008. “Ancora: Multilevel Annotated Corpora for Catalan and Spanish”, In Proceedings of the sixth international language resources and evaluation (LREC’08). Marrakech, Morocco: European Language Resources Association.

A Appendix: Example of an extractive summary

Here is an example of a text in the extractive summarisation dataset and a human-authored summary for it:

Original text

The Bank of England will tomorrow publish its annual report and accounts, which will disclose that the £319.3bn of loans made to Northern Rock is still on its books. However, the Treasury said on Friday that the Northern Rock loans would be transferred shortly to the Government's own balance sheet. The Bank's annual report for the year to March, which will be tabled in Parliament before its release to the public, is being published later than usual following a hectic few months that saw the setting-up of the Special Liquidity Scheme to ease the problems of the banking system. For now, the £350bn scheme is being treated in the accounts as an "off-balance" sheet item because the loans will be transferred to the Treasury, which is financing them with new gilts. The scheme was put together in April by the Bank's Governor, Mervyn King, and Paul Tucker, the head of the markets division, to get the banks to start lending to each other by providing them with loans in return for collateral. Although there has been some increase in lending, many of the high-street banks are still lobbying the Bank of England to put more funds in the scheme, arguing that it has not gone far enough to restore confidence. Libor, the main lending rate between banks, has stayed stubbornly high, which is still putting banks off borrowing. In return for new loans, the Bank accepts triple-A rated securitised bonds guaranteed by mortgages and credit card debt. One of the possibilities being discussed is whether the Bank would accept the extension of the scheme to include mortgages written this year; the cut-off point is currently December last year. The Bank's accounts may also show that Northern Rock is repaying its Bank of England loan faster than projected in its restructuring plan. According to Simon Ward, chief economist at fund manager New Star, the Northern Rock loan was down to £324.1bn by 31 March and has been repaid mainly by mortgage redemptions.

Summary

The Bank of England will tomorrow publish its annual report, which will disclose that the £319.3bn of loans made to Northern Rock is still on its books. However, the Treasury said on Friday that the Northern Rock loans would be transferred shortly to the Government's own balance sheet. The Bank's annual report for the year to March is being published later than usual following a hectic few months that saw the setting-up of the Special Liquidity Scheme to ease the problems of the banking system. For now, the £350bn scheme is being treated in the accounts as an "off-balance" sheet item because the loans will be transferred to the Treasury. One of the possibilities being discussed is whether the Bank would accept the extension of the scheme to include mortgages written this year; the cut-off point is currently December last year. The Bank's accounts may also show that Northern Rock is repaying its Bank of England loan faster than projected in its restructuring plan. According to Simon Ward, chief economist at fund manager New Star, the Northern Rock loan was down to £324.1bn by 31 March.