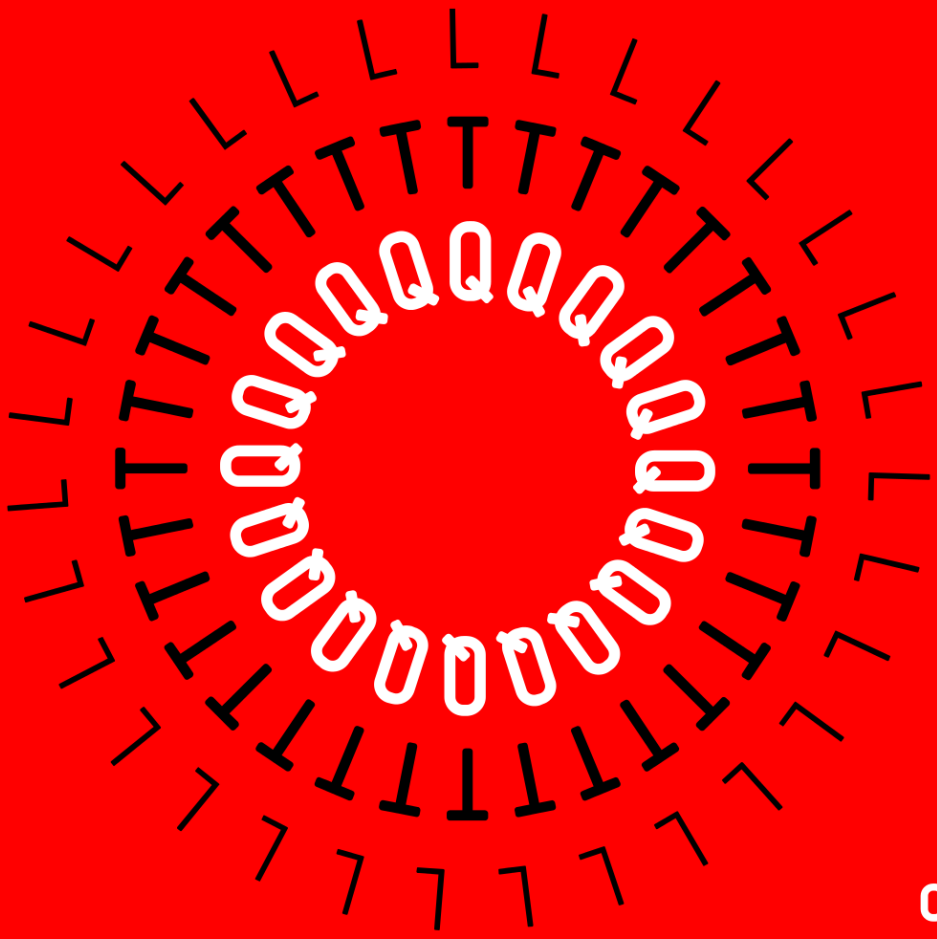


qtleap

quality translation by deep language engineering approaches

Report on first pilot version of LRTs enhanced to support deep processing

DELIVERABLE D4.7

VERSION 2. | 2014-10-28

QTLeap

Machine translation is a computational procedure that seeks to provide the translation of utterances from one language into another language.

Research and development around this grand challenge is bringing this technology to a level of maturity that already supports useful practical solutions. It permits to get at least the gist of the utterances being translated, and even to get pretty good results for some language pairs in some focused discourse domains, helping to reduce costs and to improve productivity in international businesses.

There is nevertheless still a way to go for this technology to attain a level of maturity that permits the delivery of quality translation across the board.

The goal of the QTLeap project is to research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

The deeper the processing of utterances the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches.

Deep language processing has its stepping-stone in linguistically principled methods and generalizations. It has been evolving towards supporting realistic applications, namely by embedding more data based solutions, and by exploring new types of datasets recently developed, such as parallel DeepBanks.

This progress is further supported by recent advances in terms of lexical processing. These advances have been made possible by enhanced techniques for referential and conceptual ambiguity resolution, and supported also by new types of datasets recently developed as linked open data.

The project QTLeap explores novel ways for attaining machine translation of higher quality that are opened by a new generation of increasingly sophisticated semantic datasets and by recent advances in deep language processing.

www.qtleap.eu

Funded by

QTLeap is funded by the 7th Framework Programme of the European Commission.

Supported by

And supported by the participating institutions:

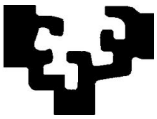
Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin



University of Basque Country

University of Groningen

Higher Functions, Lda

Revision history

Version	Date	Authors	Organisation	Description
1.0	Oct 06, 2014	Petya Osenova	IICT-BAS	First draft
1.1	Oct 21, 2014	João Silva	FCUL	Integrated text
1.2	Oct 22, 2014	Aljoscha Burchardt	DFKI	Integrated text
1.3	Oct 28, 2014	Gorka Labaka, Eneko Agirre, Nora Aranberri;	UPV-EHU	Integrated text
1.4	Oct 28, 2014	Dieke Oele, Gertjan van Noord	UG	Integrated text
1.5	Oct 28, 2014	Petya Osenova	IICT-BAS	Editing
2.0	Oct 31, 2014	Markus Egg	UBER	Review comments incorporated

Statement of originality
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Report on first pilot version of LRTs enhanced to support deep processing

DOCUMENT QTLEAP-2014-D4.7
EC FP7 PROJECT #610516

DELIVERABLE D4.7

completition
FINAL
status
SUBMITTED
dissemination level
PUBLIC

responsible
Kiril Simov (WP4 Coordinator)
reviewers
Markus Egg, Kostadin Cholakov
contributing partners
FCUL, DFKI, IICT-BAS, UBER, UPV-EHU, UG

authors
Petya Osenova, João Silva, Aljoscha Burchardt, Gorka Labaka, Dieke Oele, Markus Egg

Contents

List of Abbreviations

CoNLL	Conference on Natural Language Learning
ERG	English Resource Grammar
HPSG	Head-driven Phrase Structure Grammar
LRTs	Language Resources and Tools
MT	Machine Translation
MWE	Multiword Expressions
NLP	Natural Language Processing
POS	Part-Of-Speech
PDT	Prague Dependency Treebank
SRL	Semantic Role Labeling

1 Introduction

This deliverable describes the language resources to support the deep language processing that are provided within the deliverable D4.6 "First pilot version of language resources and tools (LRTs) enhanced to support robust deep processing". The resources for each language are uploaded to the QTLeap repository. The language resources are of two types:

- **Treebanks.** Syntactically annotated corpora to be used for the training of deep language processing tools and deep (tree-based) machine translation models.
- **Lexicons.** Dictionaries that provide semantic and valency information to support deep language processing.

The resources provided within D4.6 were described in deliverable D1.3 "Language resources and tools (LRTs) management plan". The types, size and number of resources per language were influenced by the different pre-project availability of appropriate data and the degree to which specific languages have been the object of prior research and compilation of linguistic resources.

The rest of the deliverable is organized as follows: first the treebanks due and delivered in D4.6 are presented in Chapter 2; then the lexicons due and delivered in D4.6 are described.

2 Treebanks

Treebanks are syntactically annotated resources. Deepbanks add the level of semantics on top of it. It this way the translation models are enhanced in both treebank types: monolingual and parallel.

2.1 Basque

This resource is part of Deliverable 4.6 of the QTLeap FP7 project. In its current development (15% of the intended goal of the project), it consists of 150 sentences (1,416 English tokens and 1,275 Basque tokens). The sentences are excerpts from journalistic text from the Wall Street Journal that have been manually translated into Basque to generate a parallel corpus.

The English sentences are part of the Penn Treebank corpus, and have been selected as they are already part of an English-Spanish parallel corpus[1]. In this way, we will additionally have access to a trilingual parallel treebank (English-Spanish-Basque).

The selected English sentences were manually translated, and their Basque counterparts analyzed using automatic tools. This analysis includes several levels of linguistic information for each sentence, including lemmatization and morphological analysis as well as dependency parsing trees. After the automatic analysis, a human correction phase was performed.

For English, Stanford dependency tags are used[2], whereas Basque syntactic annotation follows BDT guidelines [Aldezabal et al., 2009]. It is important to notice that both tagging styles are already included in HamleDT [Rosa et al., 2014]. Therefore, harmonization rules

[1] http://repositori.upf.edu/handle/10230/20049

[2] For a detailed description see http://nlp.stanford.edu/software/dependencies_manual.pdf

have already been developed and can be used to convert the current resource's analyses into harmonized parses, if needed.

The main motivation behind the creation of this resource was to build a high quality data set with rich grammatical information that could support the development of linguistically-informed translation tools. This treebank can be used both to guide the development of the linguistic analyzers that will be used in translation or to train, in combination with automatically annotated texts, statistical transfer module that will transform source language parses into target language ones.

2.2 Bulgarian

2.2.1 BulTreeBank-DP

BulTreeBank-DP is a monolingual dependency CoNLL-based conversion of the original HPSG-based treebank. The treebank consists of 60 % newspaper texts, 30 % literary texts and 10 % administrative and other texts. It comprises 11,900 sentences, since the sentences with ellipses have been left out. The resource complies with the annotation scheme as well as the input requirements, defined for the CoNLL contest on Dependency Parsing in 2006. The treebank includes the following levels of linguistic information: tokenization, POS, morphosyntactic features, dependency relations, coreferences.

The resource is very useful for the creation of adequate language models, which are to be used for the Bulgarian part in translation processes.

2.2.2 BulEngTreebank

This resource is part of Deliverable 4.6. It contains 920 sentences (9308 tokens) which are part of Bulgarian English Parallel Treebank. It includes English sentences from datasets distributed with the English Resource Grammar (ERG), whose domain is tourism. These sentences have already been analysed using ERG, and manually disambiguated. The sentences were translated into Bulgarian by professional translators.

Bulgarian and English sentences are aligned manually on the word level. Then they were annotated morphologically and parsed by a dependency parser. The result was manually corrected. Within the project this treebank will be extended with more pairs of sentences. Also the Bulgarian part will be represented in other formats: Universal Dependency Tagset, Minimal Recursion Semantics.

2.2.3 ParDeepBankBG

This resource is part of Deliverable 4.6. It consists of 838 sentences (21 949 tokens) from the Bulgarian English Parallel Deepbank. It includes English sentences from the English Deepbank, whose domain is journalism (Wall Street Journal). These sentences have already been analysed using ERG, and manually disambiguated. The sentences were translated into Bulgarian by professional translators. Bulgarian and English sentences are aligned manually on the word level. Then they were annotated morphologically and parsed by a dependency parser. Then the result was partially manually corrected. The dependency analyses are represented in CoNLL 2006 format. Within the project this treebank will be extended with more pairs of sentences. Also the Bulgarian part will be represented in other formats: Universal Dependency Tagset, Minimal Recursion Semantics.

2.3 Dutch

This monolingual resource contains 3000 sentences. They are parsed with the Alpino parser for Dutch and converted to Treex a-trees that are used as input for the Treex MT system. The sentences where taken from the Dutch part of the parallel OPUS-KDE4corpus, the manual of KDE which is a Windowing Manager and Graphical User Interface for the UNIX operating system. These conversion will ultimately be used for the translation of Dutch sentences within the Treex pipeline as it is planned for Pilot 1.

2.4 German

The corpus currently contains the first batch of 4600 sentences taken from the German TIGER treebank[3] that have been parsed using the Cheetah grammar for German Cramer [2011] and the PET parser. The corpus consists of files containing Trees and MRSs. STTS tags from the original TIGER corpus are preserved in the Derivation Tree. The corpus contains 40.000 tokens (4.6K, size 17 MB).

It is planned to also experiment with an alternative German HPSG grammar and compare the analyses w.r.t. to the need of the project. If the results are suitable, they will be added to the project repository too. Manual editing and selection will only be performed if relevant for the MT development within the project.

2.5 Portuguese

This resource is part of Deliverable 4.6. It is composed of 3,134 sentences (36,566 tokens) which are part of CINTIL-DeepBank (available in the META-SHARE repository). The sentences are excerpts from journalistic text from CETEMPúblico.

It includes several levels of information for each sentence, including its derivation tree obtained during parsing, its syntactic constituency tree, different renderings of MRS based representations of its meaning Copestake et al. [2005], and its fully-fledged grammatical representation in AVM format.

This is the result of a semi-automatic annotation process by means of automatic analysis by the grammar followed by a double-blind annotation followed by adjudication (see Branco and Costa [2008]), for a full description of the process).

The main motivation behind the creation of this resource was to build a high quality data set with rich grammatical information that could support the development of a large set of high level language resources and processing tools for Portuguese.

3 Lexicons

Lexicons are rich lexical databases, which include various information, such as WordNet synsets, valence frames, ontological classes. etc.

3.1 Bulgarian

3.1.1 Bulgarian Ontology-Based Lexicon

The Bulgarian Ontology-based Lexicon is organized in synsets as WordNets, but the relations between the synsets are represented via mapping to different semantic resources.

[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

The goal is they to be mapped to an appropriate ontology. In the version provided here the mapping is to the English Princeton WordNet 3.0 (WN3.0). Other mappings exist to DOLCE ontology done via OntoWordNet and via WN3.0 to SUMO ontology and other semantic resources to be presented in D5.4. P11

This version is freely available via Open Multilingual Wordnet[4].

3.1.2 Bulgarian Valency Frame Lexicon

The Valency Lexicon is a treebank-driven resource of extracted valency frames from BulTreeBank. The frames were manually curated. At the moment it comprises more than 1000 verb frames. The frames follow the surface representation in the sentences. The frame roles (or participants in the corresponding event) were assigned ontological constraints from the SIMPLE ontology (translated into Bulgarian), such as ARTEFACT, COGNITIVE FACT, etc. The representation of an entry is as follows:

<FD>
<lemma></lemma>
<def></def>
<F><F>
</FD>
where FD = Frame Description, lemma = lemma, def = definition, and F = Frame.

3.2 Portuguese

This resource is part of Deliverable 4.6. It comprises 600 lexicon entries used in LXGram, an HPSG computational grammar for deep linguistic processing of Portuguese. The lexicon was built manually. Each lexical entry is associated with a deep lexical type which is part of the type hierarchy defined in the grammar (the types associated with the 600 lexicon entries are also included in the deliverable). The deep lexical type encodes a great deal of information about the grammatical behavior of the word, such as its part-of-speech, subcategorization (valence) frame, the pattern for forming anticausative alternations, whether a verb is a raising verb or not, etc.

4 Conclusions

In this deliverable we describe two type of language resources to support the deep language processing in the project: treebanks and lexicons. The actual resources for the different languages are provided within the deliverable D4.6.

References

Izaskun Aldezabal, Maria Jesus Aranzabe, Jose Mari Arriola, and Arantza Diaz de Ilarraza. Syntactic annotation in the reference corpus for the processing of basque (epec): Theoretical and practical issues. *Corpus Linguistics and Linguistic Theory*, 5(2):241–269, 2009.

[4] http://compling.hss.ntu.edu.sg

Antonia Branco and Francisco Costa. LXGram in the Shared Task "Comparing Semantic Representations" of STEP 2008. In Johan Bos and Rodolfo Delmonte, editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 299–314. College Publications, 2008. URL http://www.aclweb.org/anthology/W08-2224. P12

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan Sag. Minimal recursion semantics: An introduction. *Research on Language & Computation*, 3(4):281–332, 2005.

Bart Cramer. *Improving the feasibility of precision-oriented HPSG parsing*. PhD thesis, Universit, 2011.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtský. HamleDT 2.0: Thirty dependency treebanks stanfordized. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, and Joseph Mariani, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341, Reykjavík, Iceland, 2014. European Language Resources Association. ISBN 978-2-9517408-8-4.