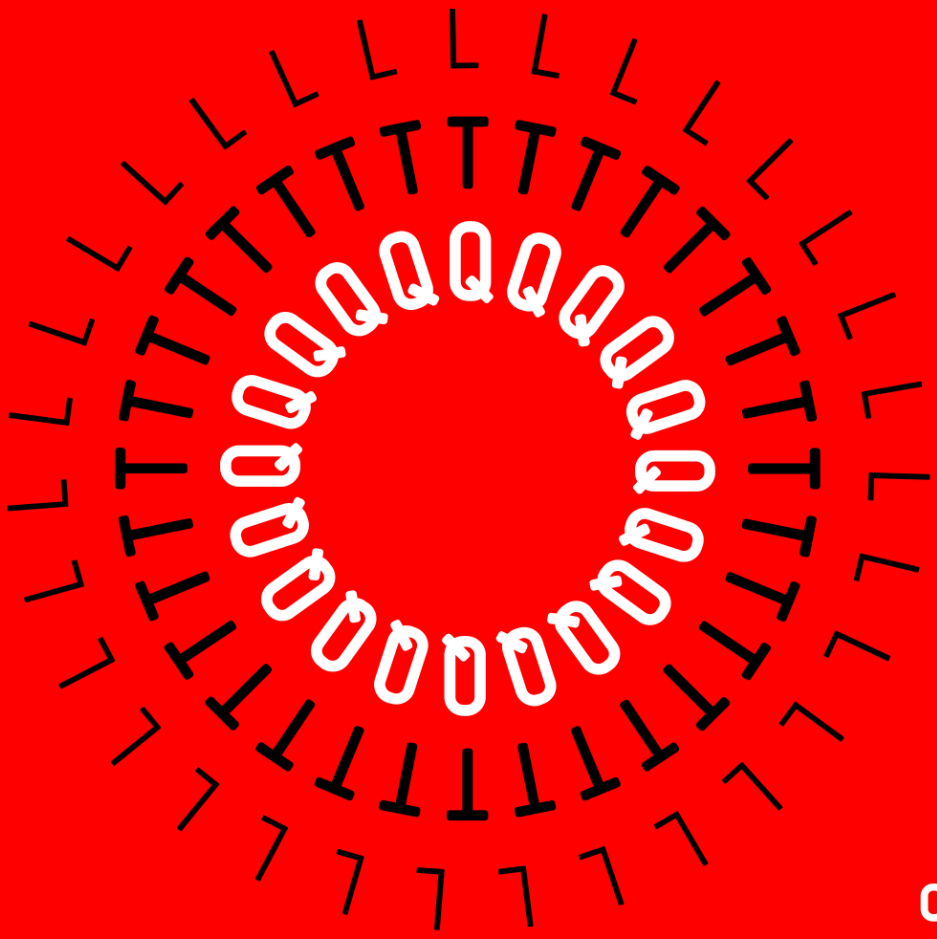


qtleap
quality translation by deep language engineering approaches

Report on pilot version of LRTs enhanced to support advanced crosslingual ambiguity resolution

DELIVERABLE D5.6

VERSION 1.4 | 2015-05-12

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:

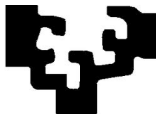
Faculty of Sciences, University of Lisbon

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
0.1	Apr 11, 2015	Eneko Agirre, Arantxa Otegi	UPV/EHU	First draft
0.2	Apr 14, 2015	Roman Sudarikov	CUNI	Sections 3.3, 5.3 and 6.3
0.3	Apr 14, 2015	Kiril Simov	IICT-BAS	Section 6.2
0.4	Apr 16, 2015	Steve Neale, João Silva	FCUL	Sections 5.5 and 6.4
0.5	Apr 16, 2015	Arantxa Otegi	UPV/EHU	Section 5.3
0.6	Apr 24, 2015	Roman Sudarikov	CUNI	Section 3.3, 5.3 and 6.3
0.7	Apr 24, 2015	Eneko Agirre	UPV/EHU	Review and harmonize
0.8	Apr 28, 2015	Roman Sudarikov	CUNI	Section 6.3
0.9	Apr 30, 2015	Eneko Agirre	UPV/EHU	Address Internal Review
1.0	May 1, 2015	Eneko Agirre	UPV/EHU	Submitted for final internal review
1.1	May 5, 2015	Kiril Simov	IICT-BAS	Address Final Internal Review
1.2	May 6, 2015	Steve Neale	FCUL	Address Final Internal Review
1.3	May 12, 2015	Roman Sudarikov, Michal Novák	CUNI	Address Final Internal Review
1.4	May 12, 2015	Eneko Agirre	CUNI	Address Final Internal Review

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 pilot version of LRTs enhanced to support advanced crosslingual ambiguity resolution

DOCUMENT QTLEAP-2015-D5.6
EC FP7 PROJECT #610516

DELIVERABLE D5.6

completition
FINAL
status
SUBMITTED
dissemination level
PUBLIC

responsible
Eneko Agirre (WP5 Coordinator)
reviewer
Gertjan van Noord
contributing partners
UPV/EHU, FCUL, CUNI, IICT-BAS

authors
Eneko Agirre, Steve Neale, Michal Novák, Arantxa Otegi, João Silva, Kiril Simov, Roman Sudarikov

Contents

1 Executive summary

The goal of the QTLeap project is to develop Machine Translation (MT) technology that goes beyond the state of the art in terms of "depth" of the methods and knowledge used. The goal of WP5 is to enhance MT with advanced crosslingual methods for the resolution of referential and lexical ambiguity by pursuing the following objectives:

1. to provide for the assembling and curation of the language resources and tools (LRTs) available to support the resolution of referential and lexical ambiguity (Task 5.1, starting M1);

2. to leverage the resolution of referential and lexical ambiguity by means of advanced crosslingual named entity and word sense resolution methods (Task 5.2, starting M1);

3. to proceed with the intrinsic evaluation of the solutions found in the previous task (Task 5.3, starting M10);

4. to contribute for high quality machine translation by using semantic linking and resolving to improve MT (Task 5.4, starting M17). In particular Pilot 2 (M24) will be devoted to check the contribution of the tools in this WP to MT.

The work reported on this document has been carried out along the plans and is based on the project Description of Work, Deliverable 1.3 ("Management plan for language resources and tools"), Deliverable 1.7 (1.7 "Language Resources and Tools Interim Report and Plan Update") and Deliverable 5.1 ("State of the art").

The present deliverable documents the language resources and tools that compose deliverable D5.5 "Pilot version of language resources and tools (LRTs) enhanced to support advanced crosslingual ambiguity resolution".

Deliverables D1.3 and D1.7 describe the new resources and tools in deliverable D5.5, as follows:

- Sense annotated corpora, for all languages in WP5 (BG Bulgarian, CS Czech, EN English, ES Spanish, EU Basque, PT Portuguese): 500K tokens aligned, 5M tokens comparable.

For easier use in the project, D5.5 actually includes the new LRTs, plus all LRTs already in D5.3 (which is described in the accompanying report, D5.4). The present report (D5.6) thus, should be read alongside D5.4.[1]

A few of the LRTs in D5.5 may have less wide distribution, but the large majority are publicly available, as described in detail in each Section below and summarized in Appendix A. For project internal purposes and the sake of replicability, all LRTs, private and public, are stored in our internal repository.

Note that English, Spanish and Bulgarian were selected to perform initial development, aimed at preparing the subsequent handling of LRTs for all the remaining languages in WP5. Thus the rest of the languages in WP5 (Basque, Czech and Portuguese) prepared Word Sense Disambiguation (WSD), Named Entity Disambiguation (NED) and coreference (CR) tools by M16. In addition, WSD and NED tools for all languages had to provide crosslingual ambiguity resolution. We thus also report in this deliverable the following:

[1] Note that D5.4 is due for resubmission on June 31, 2015

- WSD, NED and CR tools for Basque, Czech and Portuguese
- Crosslingual ambiguity resolution for all languages in WP5 (WSD, NED)

The present deliverable was preceded by D5.4 and will be followed by D5.9 (due M30). It extends D5.4 by sense-annotating further corpora, by describing tools for NED, WSD and CR in Basque, Czech and Portuguese, and by exploring crosslingual ambiguity resolution. D5.9 will report on the final versions of the language resources and tools of WP5.

Note that WP5 comprises other activities in the second year of the project that are out of the scope of this deliverable:

- Task 5.3: Intrinsic evaluation of NERC/NED, WSD, CR for Basque, Czech and Portuguese, due in M23
- Task 5.4: Experiments 5.4.1, 5.4.2 and 5.4.3, which explore several alternatives to improve MT using LRTs from WP5, due in M18 and M23.

The evaluation in Task T5.3 will be reported in D5.9 and the experiments in Task 5.4 will be reported in D5.7 ("Interim report on MT improved with offline semantic linking and resolving").

2 Introduction

The goal of the QTLeap project is to develop Machine Translation (MT) technology that goes beyond the state of the art in terms of "depth" of the methods and knowledge used. The goal of WP5 is to enhance MT with advanced crosslingual methods for the resolution of referential and lexical ambiguity by pursuing the following objectives:

1. to provide for the assembling and curation of the language resources and tools (LRTs) available to support the resolution of referential and lexical ambiguity (Task 5.1);

2. to leverage the resolution of referential and lexical ambiguity by means of advanced crosslingual named entity and word sense resolution methods (Task 5.2);

3. to proceed with the intrinsic evaluation of the solutions found in the previous task (Task 5.3);

4. to contribute for high quality machine translation by using semantic linking and resolving to improve MT (Task 5.4). In particular Pilot 2 will be devoted to check the contribution of the tools in this WP to MT.

This deliverable reports the language resources and tools (LRTs) which have been developed to contribute to high quality machine translation (Task 5.4 and Pilot 2), enabling the exploration of advanced semantic processing for machine translation. In particular, this deliverable documents the new language resources and tools (LRTs) for 6 languages (BG Bulgarian, CS Czech, EN English, ES Spanish, EU Basque, PT Portuguese) that compose deliverable D5.5 "Pilot version of LRTs enhanced to support crosslingual ambiguity resolution". These LRTs are described in Appendix A, which summarizes the new resources since the previous release, under deliverable D5.3 (resources in D5.3 are described in D5.4).

Deliverables D1.3 "Language resources and tools (LRTs) management plan" and D1.7 "Language Resources and Tools Interim Report and Plan Update" describe the resources and tools that belong to deliverable D5.5. We also report here Word Sense Disambiguation (WSD), Named Entity Disambiguation (NED) and coreference (CR) tools for Basque, Czech and Portuguese (note that the c sponding tools for Bulgarian, English and Spanish were reported in D5.4), as well as the strategy for crosslingual ambiguity resolution for all languages.

Given that this deliverable complements D5.4, and will be followed by D5.9, we decided to keep the same section numbering, for easier cross-reference. The contents are organized as follows:

- Section 5: WSD, NED and CR tools for Basque, Czech and Portuguese.

- Section 5.6: Strategy for crosslingual ambiguity resolution for all languages (WSD, NED).

- Section 6: Description of the corpora which we annotated with word senses and in addition with all available WP5 tools.

- Appendix A: Summary of LRTs described in this deliverable, alongside availability information.

5 NED, WSD and Coreference tools

5.3 Basque

5.3.1 NED

The ixa-pipe-ned-ukb module performs the Named Entity Disambiguation (NED) task based on UKB, a graph-based Word Sense Disambiguation (WSD) tool (see next section). In this case, the Wikipedia graph built from the hyperlinks between Wikipedia articles is used for the processing. This tool was successfully used for English NED (Agirre et al. [2015]).

The input of the module is text where named entity mentions have been recognized and represented using the Natural Language Processing Annotation Format (NAF)(Fokkens et al. [2014])[2]. In the output it returns the corresponding Basque Wikipedia in NAF format.

The tool is released under license GPLv3.0[3]. The tool is partly funded by QTLeap, as the wrapper to read and produce NAF has been developed in this project.

5.3.2 WSD

UKB is a collection of programs for performing graph-based Word Sense Disambiguation[4]. It applies the so-called Personalized PageRank on a Knowledge Base (KB) to rank the vertices of the KB and thus perform disambiguation. We used WordNet 3.0 as the KB for performing WSD.

ixa-pipe-wsd-ukb takes lemmatized and PoS tagged text in NAF format as standard input and outputs NAF. The tool is released under license GPLv3.0, packaged with the resources to run it on Basque[5]. The tool has been developed independently from QTLeap.

5.3.3 Coreference

The module of Basque coreference resolution (ixa-pipe-coref-eu) is an adaptation of the Stanford Deterministic Coreference Resolution (Lee et al. [2013]), which gives state-of-the-art performance for English. The original system applies a succession of ten independent deterministic coreference models or sieves. During the adaptation process, firstly, a baseline system has been created which receives as input texts processed by Basque analysis tools and uses specifically adapted static lists to identify language dependent features like gender, animacy or number. Afterwards, improvements over the baseline system have been applied, adapting and replacing some of the original sieves, taking into account that morpho-syntactic features are crucial in the design of the sieves for agglutinative languages like Basque.

The module needs a NAF document annotated with lemmas, entities and constituents, and outputs a NAF document.

The tool is released under license GPLv3.0[6]. The tool is partly funded by QTLeap, as the wrapper to read and produce NAF has been developed in this project.

[2] http://wordpress.let.vupr.nl/naf/
https://github.com/newsreader/NAF/blob/master/naf.pdf?raw=true

[3] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-ned-ukb.tar.gz

[4] https://github.com/asoroa/naf_ukb

[5] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-wsd-ukb.tar.gz

[6] http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-coref-eu.tar.gz

5.4 Czech

All the tools mentioned below produce results in Treex and NAF formats.

5.4.1 NED

Currently, there is no publicly available implementation for NED of Czech. During the preparation phase for Named Entities Disambiguation task we created the Named Entities Linking table. Each row of that table consisted of the lemmatized Czech Wikipedia article's title, Czech Wikipedia URL and English DBpedia URL. In order to make this table we downloaded 2 dumps: Czech Wikipedia dump (containing Czech titles and corresponding Wikipedia URLs), English-Czech DBpedia dump (containing Czech labels and English DBpedia URLs). Czech labels from the DBpedia were mapped to the titles of corresponding Czech Wikipedia articles thus creating the resulting table. We additionally applied lemmatization and tagging for each title using MorphoDiTa (Straková et al. [2014]).

Named Entity Disambiguation was done in two steps. During the first step we used the Treex block for the NameTag tool (Straková et al. [2014]) to detect named entities in the corpus. During the second step we used the previously created Named Entities Linking table in the following way: for each entity that was detected by NameTag we lemmatized its form and then searched the table for the occurrences of this lemmatized form. We used lemmatization to resolve the problem of forms' inflection. If the search returned results, we looked for the DBpedia URL and labeled the entity if we could find one. In case of ambiguity the algorithm picked up "most popular" article. The popularity of the article was computed using Wikipedia page-to-page link records, so the article with the highest number of reference links was preferred. We are looking forward to further improving of the algorithm by adding the context from the Wikipedia articles. These improvements will be evaluated and reported in the next deliverable in WP5.

The development of Treex wrappers for MorphoDiTa and NameTag is partly funded by the project. They are available under open-source license (Perl Artistic + GPL) at GitHub repository.[7] The tools (MorphoDiTa and NameTag) themselves are also open source (LGPL) and available from GitHub or `http://www.lindat.cz/`.

5.4.2 WSD

Experiments in Czech WSD [Honetschläger, 2003, Semecký, 2007, Hajič et al., 2009] typically use the Prague Dependency Treebank (PDT) [Hajič et al., 2006, Bejček et al., 2012], which provides valency frame reference annotation, i.e., word sense labeling for all verbs and many other content words. PDT word senses are based on the PDT-Vallex Czech valency lexicon [Hajič et al., 2003, Urešová, 2011]. A mapping [Urešová et al., 2014][8] connects it to the EngVallex valency lexicon [Cinková, 2006], which itself contains links to PropBank and can thus be mapped to English WordNet [Pazienza et al., 2006]. Note that PDT word senses do not form a hierarchy, which makes it incompatible with graph-based WSD (see Sections 5.3.2 and 5.5.2).

Although there is a WordNet for Czech [Pala et al., 2011], it is typically not used for WSD tasks. It is based on an outdated version of the Princeton WordNet (2.0) and it has been further modified, and so its mapping onto current English WordNet is not trivial.

[7] `https://github.com/ufal/treex/`
[8] `http://ufal.mff.cuni.cz/czengvallex`

The Czech WSD annotation developed herein uses two approaches: First, a tool based on [Dušek et al., 2014], which uses the VowpalWabbit linear classifier on top of automatic deep syntactic analysis and achieves high performance for verbal WSD on PDT data. This is used for real user scenarios.[9] Second, for the WSD-annotated parallel corpus, we opted for a more straightforward way of achieving compatibility with English WordNet IDs: since the corpus contains the same sentences as the EN-ES parallel corpus provided in D5.3, we could use the English WordNet ID annotation from this corpus and project it onto Czech words using GIZA++ word alignment. This method will be evaluated for the next deliverable in WP5. In case the evaluation is disappointing, we plan to improve the current approach.

5.4.3 Coreference

The coreference resolution system for Czech consists of multiple modules, each of them aiming at a specific type of coreference: coreference of reflexive pronouns, relative pronouns, zeros, personal and possessive pronouns in 3rd person and coreference of noun phrases. Coreference relations are annotated between the nodes of dependency trees that serve as a deep syntax representation of sentences. This enables the system to take advantage of rich linguistic annotations available in the trees as well as to resolve coreference even for subject pronouns dropped from the surface representation (zeros), which is a common practice in Czech.

Due to the pro-drop nature of Czech, the places where a subject is unexpressed have to be identified before proceeding to coreference resolution of zeros. This is performed based on the syntactic information and a special node representing the zero is added to the deep syntax tree. The grammatical categories of the newly added zero are then inferred from the grammatical categories of its governing verb. Reconstruction of zeros is implemented in the Treex block `A2T::CS::AddPersPron`.

The modules targeting coreference of *relative and reflexive pronouns* are based on the rules presented in (Nguy [2006]). The rules exploit morphological information together with syntactic structure and stick to the principle that the antecedent of a reflexive pronoun is the sentence's subject whereas the antecedent of the relative pronoun usually directly governs the relative clause introduced by the pronoun. These resolvers are implemented in Treex blocks `A2T::CS::MarkRelClauseCoref` and `A2T::CS::MarkReflpronCoref`.

Unlike the previous cases, resolution of *personal and possessive pronouns and zeros in 3rd person* is treated by a machine learning approach. It adheres to a so-called mention-ranking model (Denis and Baldridge [2007]) with features capturing the distance between the mentions (in words, clauses and sentences), grammatical information (e.g., agreement in their numbers and genders) as well as semantic information (semantic roles, classes in the Czech part of EuroWordNet (Vossen [1998])). The currently used model is built using logistic regression in Vowpal Wabbit[10] and is available in the Treex block `A2T::CS::MarkTextPronCoref`. A more detailed description and evaluation can be found in (Nguy et al. [2009], Bojar et al. [2012]).

Coreference of *noun phrases* is modeled in the same way as the pronouns and zeros in the previous case. However, the feature set is more oriented on lexical features (equality of

[9] The tool has been developed independently of QTLeap is implemented as a Treex block `A2T::SetValencyFrameRefVW` and is available under Perl Artistic and GPL license in the Treex Git repository: `https://github.com/ufal/treex/`.

[10] `https://github.com/JohnLangford/vowpal_wabbit`

head lemmas), semantic features (synonymy approximation extracted from the English-Czech parallel corpus CzEng 0.9 (Bojar et al. [2009]), EuroWordNet classes) and the information about named entities. Unlike the previous modules, the module for noun phrases does not have a Treex binding, yet. A more detailed description and evaluation can be found in (Novák and Žabokrtský [2011]).

All modules are available under open-source license (Perl Artistic + GPL) and the Treex blocks can be downloaded from its Git repository.[11]

5.5 Portuguese

5.5.1 NED

The named entity disambiguation pipeline for Portuguese uses DBpedia Spotlight (Daiber et al. [2013]) to find links to resources about entities identified in pre-processed input text. It creates a process to run a Portuguese extraction of DBpedia Spotlight on a local server, then takes an input text pre-processed with lemmas, Part of Speech tags and named entities using the LX-Suite (Branco and Ricardo Silva [2006]) and converts it to the 'spotted' format understood by Spotlight. This spotted input text is then disambiguated using DBpedia Spotlight, returning among other information links to existing Portuguese DBpedia resource pages for each named entity discovered.

For each Portuguese DBpedia resource page link found, the tool performs a DBpedia sparql query to find any English words that the link in question relates back to. These results can then be used to determine the corresponding English DBpedia resource page link, for example: http://pt.dbpedia.org/resource/Paquistão relates to 'Pakistan', thus the equivalent link in English must be http://dbpedia.org/resource/Pakistan. This process has been found to return working English resource links in almost all cases, with the exception of Portuguese resource links that despite existing contain no actual information (having perhaps been corrupted, or created and then for some reason deleted later).

The output displays each potential named entity found in the input text with: its positional offsets (sentence and position within sentence); the disambiguated Portuguese DBpedia resource link (if found); and the corresponding English DBpedia resource link (if found).

5.5.2 WSD

For WSD, a pipeline was used that takes pre-processed input text and runs it through the UKB word-sense disambiguation algorithm (Agirre and Soroa [2009]). The pre-processed texts, .txt files lemmatized and PoS-tagged using the LX-Suite, are passed as an argument to the pipeline, which converts the text to the context format recognized by UKB. The Lexical Knowledge Base (LKB) from which UKB returns word senses within the pipeline has been generated from an extraction of the Portuguese MultiWordNet.[12]

The output displays each potentially ambiguous word (noun, verb, adjective or adverb) found in the input text with: incrementing ID numbers; its UKB context (sentence number); its UKB word id (position within sentence); its part-of-speech; its lemma; whether or not it was tagged by UKB; the Portuguese MultiWordNet sense returned by UKB; and the ILI code.

[11] `https://github.com/ufal/treex/`

[12] The MultiWordNet project: `http://multiwordnet.fbk.eu/english/home.php`. Accessed 2015-04-09.

5.5.3 Coreference

As an initial study for the coreference pipeline, a decision tree classifier was experimented with. Given a pair of expressions, the classifier returns a true or false value that indicates whether those expressions are coreferent. The classifier was trained over the Summ-it Corpus (Collovini et al. [2007]) using the J48 algorithm in the Weka machine-learning toolkit (Hall et al. [2009]). The most relevant features, according to the work of (de Souza et al. [2008]), were extracted from Summ-It and used to train the J48 algorithm, with default parameters. The extracted features were the following:

- cores-match, which indicates whether the "cores" of the two expressions (i.e. their heads) have the same form;
- gender-agreement, which indicates whether the cores of the two expressions have matching gender;
- number-agreement, which is similar to the previous one, but for number;
- distance, which indicates the distance, in sentences, between the two expressions (if the two expressions occur in the same sentence, their distance is zero);
- antecedent-is-pron, which indicates whether the antecedent (the first markable) is a pronoun;
- anaphora-is-pron, which indicates whether the anaphora (the second markable) is a pronoun;
- both-proper-names, which indicates whether both markables are proper names;

The resulting decision tree produced by J48 turned up to very simple and boils down to comparing the cores and the morphological information (gender and number) of the two expressions. As such, we found it easier to directly implement equivalent tests in-code instead of having to feed the extracted features to the Weka J48 classifier proper.

The script for resolving coreference runs over text that has been parsed by a constituency parser and by a dependency parser. Constituency information allows to find spans of text that correspond to noun phrases, while the dependency information is used when finding the heads of constituents. Sets of markables are gathered, formed by all the noun phrases and pronouns in the document. For every pair of markables, the relevant features are extracted and relevant tests are performed, returning a true or false value for each pair of markables.

The script[13] is licensed under an Apache License 2.0.

5.6 Crosslingual ambiguity resolution

Assuming that concepts and instances are shared across languages and cultures, at least to a great extent, it is in theory possible to construct a common repository of concepts and instances.

Following our design for aligned ontologies (D5.4, Section 3), WSD and NED tools return, respectively interlingual concept identifiers and instance identifiers, as follows:

[13] http://nlx-server.di.fc.ul.pt/~jsilva/Tool-Coref-PT.tar.gz

- The interlingual concept id's are inspired in EuroWordNet Vossen [1998], which presented the design of a multilingual database with wordnets for several languages. The design was based on the ILI, based on the English wordnet. Via this index, the languages are interconnected at the senses level, so that it is possible to go from the words in one language to similar words in any other language via equivalent senses. Current ILIs are based on the English WordNet 3.0 synset numbers, and are strings like the following: `ili-30-05799212-n`, where `30` stands for the 3.0 WordNet version, `05799212` corresponds to the English WordNet 3.0 synset number, and `-n` to the PoS of the synset. Note that the synset number in the ILI has 9 digits, which are obtained appending 0 to the 8 digits of the WordNet 3.0 synsets. This allows some room to incorporate concepts which are not found in the English WordNet, although, given the fact that all translations involve English, this possibility is not needed in QTLeap.

- Instance id's are based on English DBpedia v3.9 URIs.

Producing ILI's is straightforward in the cases where the wordnets are aligned to the English version. This is possible for all the languages covered in our work.

Producing English DBpedia v3.9 URIs is also straightforward for NED tools, as DBpedia maintains interlingual links between articles in different languages.

6 Annotated corpora

This Section describes the corpora which have been annotated with the WP5 tools mentioned above. All the corpora below have been packaged in two multilingual corpora released through meta-share[14] and CLARIN Lindat[15]:

- Europarl-QTLeap WDS/NED corpus. In addition to BG, CS, EN, ES, PT subsets of the Europarl parallel corpus[16] Koehn [2005], it contains an EN-EU parallel corpus from non-Europarl sources.
- QTLeap WSD/NED corpus. It contains Batches 1 and 2 of the QTLeap corpus[17] annotated.

Note that, due to licensing restrictions, we are only allowed to redistribute part of the EN-EU dataset. The rest of the EN-EU dataset is available to project partners in the project internal repository.

The Europarl-QTLeap WSD/NED corpus is distributed under the license CC BY 4.0.

The QTLeap WSD/NED corpus is distributed under the license CC BY-NC-SA 4.0.

All the language pairs annotated parallel corpora instead of comparable corpora (as initially planned in the DoW), as this provides better quality for training machine translation systems.

6.1 Basque-English

Given that Europarl does not include Basque, we gathered publicly available and private corpora. Regarding publicly available corpora, we focus on the GNOME corpus (Tiedemann [2012]). Regarding private corpora, we have access to the translation memories of Elhuyar Foundation (obtained via Eleka, member of the Advisory Board of Potential Users), which we cannot redistribute.

Prior to being annotated with ixa-pipe-wsd-ukb, ixa-pipe-ned-ukb and ixa-pipe-coref-eu, we preprocessed the corpus with the tools described in D5.4. Thus, the annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation
- Word Sense Disambiguation
- Dependency parser
- Coreference

[14] http://metashare.metanet4u.eu/go2/europarl-qtleap-wsdned-corpus
http://metashare.metanet4u.eu/go2/qtleap-wsdned-corpus

[15] https://lindat.mff.cuni.cz/

[16] http://www.statmt.org/europarl/

[17] The QTLeap corpus is described in deliverable D2.5

These are the annotated corpora:

- Parallel corpus
  - Elhuyar-QTLeap WSD/NED corpus (private)

Corpus	EU
tokens	5,157,531
terms	5,157,531
linked to WordNet	2,214,554 (42.94%)
entities	68,087
linked to DBpedia	27,412 (40.26%)
coreference chains	724,912

Table 1: Statistics on Elhuyar-QTLeap WSD/NED corpus (Basque)

  - GNOME section of the Europarl-QTLeap WSD/NED corpus (public)

Corpus	EU
tokens	4,194,823
terms	4,194,823
linked to WordNet	1,940,424 (46.26%)
entities	45,801
linked to DBpedia	21,118 (46.11%)
coreference chains	563,570

Table 2: Statistics on the GNOME section of the Europarl-QTLeap WSD/NED corpus (Basque)

  - QTLeap WSD/NED corpus: batch 1 and 2

Corpus	EU
tokens	53,239
terms	53,239
linked to WordNet	24,691 (46.38%)
entities	869
linked to DBpedia	252 (29.00%)
coreference chains	5,542

Table 3: Statistics on QTLeap WSD/NED corpus (Basque)

6.2 Bulgarian-English

The Bulgarian corpora were processed by BTB-Pipeline (more details in deliverable D5.4). The annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagging and lemmatization
- Named Entity Recognition and Classification

- Named Entity Disambiguation
- Word Sense Disambiguation
- Dependency parsing
- Coreference

The resulting annotations are represented in NAF. The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: We used part of Bulgarian-English Europarl corpus v7.0, which intersects with Spanish-English corpus, provided in D5.3. We have got 3M tokens out of total 14M.

Corpus	BG
tokens	4,840,787
terms	4,840,787
linked to WordNet	1,282,507 (26.5%)
entities	61,218
linked to DBpedia	61,218 (100%)
coreference chains	30,998

Table 4: Statistics on annotated parallel corpora (Bulgarian)

    Together with Bulgarian-English corpus reported in deliverable D5.4 the total number of tokens is higher than 5M.
  - QTLeap WSD/NED corpus: batch 1 and 2

Corpus	BG
tokens	67,591
terms	67,591
linked to Wordnet	12,627 (18.7%)
entities	180
linked to DBpedia	180 (100%)
coreference chains	306

Table 5: Statistics on QTLeap WSD/NED corpus (Bulgarian)

6.3 Czech-English

The whole annotation process is run in Treex scenario. All processes are implemented as Treex blocks. Word sense disambiguation was based on the Valency lexicon disambiguation. PoS tagger MorphoDiTa and NERC tool NameTag annotation processes were described in deliverable D5.4. The annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation
- Word Sense Disambiguation
- Dependency parser
- Coreference

These are the annotated corpora:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: We used part of Czech-English Europarl corpus v7.0, which intersects with Spanish-English corpus, provided in D5.3. We have got 9M tokens out of total 14M.

Corpus	CZ
tokens	9,094,542
terms	9,094,542
linked to Wordnet	4,474,614(49.2%)
entities	295,871
linked to DBpedia	116,650(39.4%)
coreference chains	199,590

Table 6: Statistics on annotated Europarl parallel corpora (Czech)

- QTLeap WSD/NED corpus: batch 1 and 2

Corpus	CZ
tokens	71,061
terms	71,061
linked to Wordnet/Vallex	11,060(15.5%)
entities	1715
linked to DBpedia	572 (33.3%)
coreference chains	1027

Table 7: Statistics on QTLeap WSD/NED corpus (Czech)

6.4 Portuguese-English

We have annotated the freely available Europarl v7.0 parallel corpus (5M tokens, 160K sentences). Prior to being annotated with the NED, WSD and coreferencing pipelines, the corpus was pre-annotated using LX-Suite (tokenization, PoS, lemmatization and morphological information) and LX-NER (named entity recognition), as well as being constituency and dependency parsed for the coreferencing task. Summarizing, the Portuguese annotated corpus contain:

- Lemmas, part-of-speech tags and morphological information as part of the pre-processing provided by the LX-Suite on the raw corpus.
- Word senses, provided from the output of the WSD pipeline.
- URIs for Portuguese DBpedia links, provided by the output of the NED pipeline.
- Coreference information, provided by the output of the coreferencing pipeline.

The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: We annotated a 160 Kline subset of the Portuguese-English Europarl corpus v7.0. The intersection of this subset with the English side of the Spanish-English corpus is 91%.

Corpus	PT
tokens	5,044,533
terms	2,051,531
linked to WordNet	659,212 (32.13%)
entities	172,539
linked to DBpedia	102,804 (59.58%)
coreferent pairs	5,885,270

Table 8: Statistics on Europarl-QTLeap WSD/NED corpus (Portuguese)

  - QTLeap WSD/NED corpus: batch 1 and 2

Corpus	PT
tokens	72,018
terms	29,985
linked to WordNet	6,116 (20.40%)
entities	3,799
linked to DBpedia	1,868 (49.17%)
coreferent pairs	183

Table 9: Statistics on QTLeap WSD/NED corpus (Portuguese)

6.5 Spanish-English

The corpora reported here extends the ones released previously in D5.3. We used the same ixa-pipe tools (Agerri et al. [2014]) described in D5.4 on this larger corpora. Thus, the annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation

- Word Sense Disambiguation
- Constituent parser
- Coreference

The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: Comprises the whole Spanish-English Europarl v7.0 corpus. Note that the English side of the Spanish-English Europarl corpus was the one chosen as the main English annotated corpus.

Corpus	ES
tokens	56,991,166
terms	56,991,166
linked to WordNet	20,192,103 (35.43%)
entities	2,183,742
linked to DBpedia	1,211,663 (55.49%)
coreference chains	938,708

Table 10: Statistics on Europarl-QTLeap WSD/NED corpus (Spanish)

  - QTLeap WSD/NED corpus: batch 1 and 2

Corpus	ES
tokens	71,989
terms	71,989
linked to WordNet	22,704 (31.54%)
entities	4,313
linked to DBpedia	3,175 (73.61%)
coreference chains	705

Table 11: Statistics on QTLeap WSD/NED corpus (Spanish)

6.6 English side of parallel corpora

We used the same ixa-pipe tools (Agerri et al. [2014]) described in D5.4 on the annotation of these corpora. Thus, the annotation in each document is an output of the following annotation process:

- Tokenization
- Part of Speech tagger and lemmatization
- Named Entity Recognition and Classification
- Named Entity Disambiguation
- Word Sense Disambiguation
- Constituent parser

- Coreference

The annotated corpora are the following:

- Parallel corpus
  - Europarl-QTLeap WSD/NED corpus: The English side of the Europarl-QTLeap WSD/NED corpus contains two corpora. One is the English side of the EN-ES Europarl corpus v7.0, which aligns with the Bulgarian, Czech, Spanish and Portuguese sides of the respective Europarl corpus. The second is the Basque side of the publicly available EN-EU corpus, which is not related to Europarl.

Corpus	EN
tokens	52,244,552
terms	52,244,552
linked to WordNet	22,706,391 (43.46%)
entities	1,914,802
linked to DBpedia	1,498,128 (78.24%)
coreference chains	1,248,690

Table 12: Statistics on Europarl-QTLeap WSD/NED corpus, English side of the Spanish-English Europarl

Corpus	EN
tokens	5,329,594
terms	5,329,594
linked to WordNet	2,145,509 (40.26%)
entities	271,396
linked to DBpedia	133,223 (49.09%)
coreference chains	59,126

Table 13: Statistics on the GNOME section of the Europarl-QTLeap WSD/NED corpus, English side of the Basque-English corpus

  - QTLeap WSD/NED corpus: batch 1 and 2

Corpus	EN
tokens	68,913
terms	68,913
linked to WordNet	25,807 (37.45%)
entities	2,999
linked to DBpedia	1,950 (65.02%)
coreference chains	1,199

Table 14: Statistics on QTLeap WSD/NED corpus, English side

7 Final remarks

This deliverable reports on the LRTs curated and produced in WP5 in the period from M12 to M18, as described in D1.3 and D1.7, comprising 6 languages (BG Bulgarian, CS Czech, EN English, ES Spanish, EU Basque, PT Portuguese). It includes two multilingual corpora:

- Europarl-QTLeap WDS/NED corpus. In addition to BG, CS, EN, ES and PT subsets of the Europarl parallel corpus, it contains an EN-EU parallel corpus from non-Europarl sources.
- QTLeap WSD/NED corpus. It contains annotated versions of Batches 1 and 2 of the QTLeap corpus, including BG, CS, EN, ES, EU and PT.

Both multilingual corpora have been annotated with WSD, NED and coreference information. At this stage the goal was to provide 500K tokens from parallel corpora, and 5M tokens from comparable corpora. All languages focused on parallel corpora, as it enables better machine translation, in some cases exceeding the required sizes.

References

R. Agerri, J. Bermudez, and G. Rigau. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), may 2014.

E. Agirre and A. Soroa. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41. Association for Computational Linguistics, 2009.

E. Agirre, A. Barrena, and A. Soroa. Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation. 2015. URL `http://arxiv.org/abs/1503.01655`.

E. Bejček, J. Panevová, J. Popelka, P. Straňák, M. Ševčíková, J. Štěpánek, and Z. Žabokrtský. Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In *Proceedings of COLING 2012: Technical Papers*, Mumbai, 2012.

Ondřej Bojar, Zdeněk Žabokrtský, Miroslav Janíček, Václav Klimeš, Jana Kravalová, David Mareček, Václav Novák, Martin Popel, and Jan Ptáček. CzEng 0.9, 2009.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. The joy of parallelism with CzEng 1.0. In *Proceedings of LREC 2012*, Istanbul, Turkey, 2012. European Language Resources Association.

A. Branco and J. Ricardo Silva. A Suite of Shallow Processing Tools for Portuguese: LX-Suite. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations*, EACL '06, pages 179–182. Association for Computational Linguistics, 2006.

S. Cinková. From PropBank to EngValLex: adapting the PropBank-Lexicon to the valency theory of the functional generative description. In *Proceedings of LREC 2006, Genova, Italy*, 2006.

S. Collovini, T. I. Carbonel, J. T. Fuchs, J. C. Coelho, L. Rino, and R. Vieira. Summ-it: um corpus anotado com informações discursivas visando sumarização automática. In *Proceedings of TIL 2007*, 2007.

J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, pages 121–124. ACM, 2013.

J. G. C. de Souza, P. N. Gonçalves, and R. Vieira. Learning Coreference Resolution for Portuguese Texts. In A. Teixeira, V. de Lima, L. de Oliveira, and P. Quaresma, editors, *Computational Processing of the Portuguese Language*, volume 5190 of *Lecture Notes in Computer Science*, pages 153–162. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85979-6.

Pascal Denis and Jason Baldridge. A ranking approach to pronoun resolution. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1588–1593, 2007.

O. Dušek, J. Hajic, and Z. Uresova. Verbal Valency Frame Detection and Selection in Czech and English. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 6–11. Association for Computational Linguistics, 2014.

A. Fokkens, A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. Robert van Hage, and P. Vossen. NAF and GAF: Linking Linguistic Annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 2014.

J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, and Y. Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL-2009*, Boulder, Colorado, USA, 2009.

J. Hajič, J. Panevová, Z. Urešová, A. Bémová, V. Kolářová, and P. Pajas. PDT-VALLEX: creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The 2nd Workshop on Treebanks and Linguistic Theories*, volume 9, page 57–68, 2003.

J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková Razímová, and Z. Urešová. *Prague Dependency Treebank 2.0*. Number LDC2006T01. LDC, Philadelphia, PA, USA, 2006.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009. ISSN 1931-0145.

V. Honetschläger. Using a Czech valency lexicon for annotation support. In *Text, Speech and Dialogue*, pages 120–125. Springer, 2003.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT. URL `http://mt-archive.info/MTS-2005-Koehn.pdf`.

H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Comput. Linguist.*, 39(4):885–916, 2013. ISSN 0891-2017.

Giang Linh Nguy. Návrh souboru pravidel pro analýzu anafor v českém jazyce. Master's thesis, MFF UK, Prague, Czech Republic, 2006. In Czech.

Giang Linh Nguy, Václav Novák, and Zdeněk Žabokrtský. Comparison of classification and ranking approaches to pronominal anaphora resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, London, UK, 2009. The Association for Computational Linguistics. ISBN 978-1-932432-64-0.

Michal Novák and Zdeněk Žabokrtský. Resolving noun phrase coreference in czech. *Lecture Notes in Computer Science*, 7099:24–34, 2011. ISSN 0302-9743.

Karel Pala, Tomáš Čapek, Barbora Zajíčková, Dita Bartůšková, Kateřina Kulková, Petra Hoffmannová, Eduard Bejček, Pavel Straňák, and Jan Hajič. Czech WordNet 1.9 PDT, 2011. URL `http://hdl.handle.net/11858/00-097C-0000-0001-4880-3`.

M. Pazienza, M. Pennacchiotti, and F. Zanzotto. Mixing WordNet, VerbNet and PropBank for studying verb relations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), 2006.

J. Semecký. Verb valency frames disambiguation. *The Prague Bulletin of Mathematical Linguistics*, (88):31–52, 2007.

Jana Straková, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P14/P14-5003.pdf`.

J. Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218. European Language Resources Association (ELRA), 2012.

Z. Urešová. *Valenční slovník Pražského závislostního korpusu (PDT-Vallex)*. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Praha, Czechia, ISBN 978-80-904571-1-9, 375 pp., 2011.

Z. Urešová, O. Dušek, E. Fučíková, J. Hajič, and J. Šindlerová. Bilingual English-Czech valency lexicon linked to a parallel corpus. In *Proceedings of LAW-NAACL*, Boulder, Colorado, 2014. To appear.

Piek Vossen, editor. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. ISBN 0-7923-5295-5.

A Summary of availability

Name of LRT	language	QTLeap	License	URL
Europarl-QTLeap WDS/NED corpus	BG,CS,EN, ES,EU,PT	Yes	CC-BY v4.0	http://metashare.metanet4u.eu/go2/europarl-qtleap-wsdned-corpus
ixa-pipe-ned-ukb	EU	Yes	GPLv3.0	http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-ned-ukb.tar.gz
ixa-pipe-wsd-ukb	EU	No	GPLv3.0	http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-wsd-ukb.tar.gz
ixa-pipe-coref-eu	EU	Yes	GPLv3.0	http://ixa2.si.ehu.eus/ixa-pipes/eu/ixa-pipe-coref-eu.tar.gz
MorphoDiTa Treex wrapper	CS	Yes	GPLv3.0 + Perl Artistic	https://github.com/ufal/treex/
NameTag Treex wrapper	CS	Yes	GPLv3.0 + Perl Artistic	https://github.com/ufal/treex/
QTLeap WDS/NED corpus	BG,CS,EN, ES,EU,PT	Yes	CC-BY-NC-SA v4.0	http://metashare.metanet4u.eu/go2/qtleap-wsdned-corpus
Portuguese coreference tool	PT	Yes	Apache License 2.0	http://nlx-server.di.fc.ul.pt/~jsilva/Coref-PT.tgz

Table 15: Summary of publicly available LRTs mentioned in D5.6. QTLeap column for those LRTs which have been (partially) funded by QTLeap. QTLeap corpus are also available through CLARIN Lindat (https://lindat.mff.cuni.cz/)