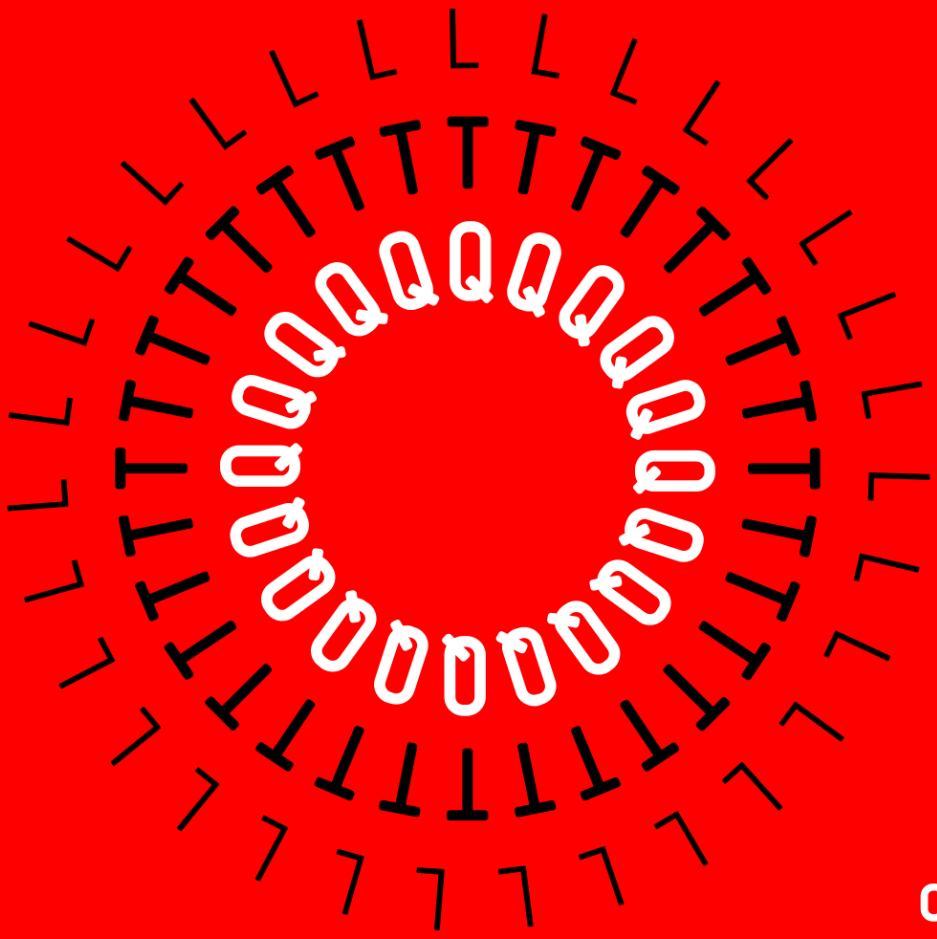


qtleap
quality translation by deep language engineering approaches

Interim report on MT improved with offline semantic linking and resolving

DELIVERABLE D5.7

VERSION 0.20 | 2015-11-16

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:

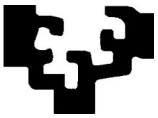
Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin



University of Groningen

Higher Functions, Lda

Revision history

Version	Date	Authors	Organisation	Description
0.1	July 1, 2015	Martin Popel	CUNI	First draft
0.2	August 30, 2015	Michal Novák	CUNI	Section on gazetteers
0.3	September 2, 2015	Eneko Agirre	UPV/EHU	Structure
0.4	September 3, 2015	Steven Neale	FCUL	Experiments 5.4.1
0.5	September 7, 2015	Aljoscha Burchardt	DFKI	Additional experiments on German
0.6	September 25, 2015	Martin Popel	CUNI	Results
0.7	September 28, 2015	Gorka Labaka	UPV/EHU	Additional experiments on word senses and NE
0.8	October 19, 2015	Arantxa Otegi, Nora Aranberri	UPV/EHU	Evaluation of Basque LRTs
0.9	October 20, 2015	Michal Novák	CUNI	Extra experiments on coreference
0.10	October 20, 2015	Steven Neale	FCUL	Evaluation of Portuguese LRTs
0.11	October 25, 2015	Michal Novák	CUNI	Revising the whole D5.7
0.12	October 26, 2015	Gorka Labaka	UPV/EHU	Additional experiments on domain corpora from Wikipedia
0.13	October 26, 2015	Steven Neale	FCUL	Revising section on Experiments 5.4.1
0.14	October 26, 2015	Rosa Del Gaudio	HF	Adding information on Wikipedia gazetteers
0.15	October 26, 2015	Gertjan van Noord	UG	Adding concepts in en→nl
0.16	October 29, 2015	Kiril Simov	IICT-BAS	Adding experiments for Bulgarian
0.17	November 1, 2015	Michal Novák	CUNI	Checking formal aspects
0.18	November 7, 2015	Will Roberts	UBER	Add multiword experiment
0.19	November 12, 2015	Kiril Simov	IICT-BAS	Last details for Bulgarian
0.20	November 12, 2015	Eneko Agirre, Martin Popel	UPV/EHU, CUNI	Reaction to internal review, final editing

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.

Interim report on MT improved with offline semantic linking and resolving

DOCUMENT QTLEAP-2015-D5.7
EC FP7 PROJECT #610516

DELIVERABLE D5.7

completition
FINAL
status
SUBMITTED
dissemination level
PUBLIC

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

authors
Martin Popel, Michal Novák, Eneko Agirre, Nora Aranberri, Gorka Labaka, Arantxa Otegi, Roman Sudarikov, Steven Neale, João Silva, António Branco, Aljoscha Burchardt, Kiril Simov, Gertjan van Noord, Rosa del Gaudio, Will Roberts, Markus Egg

Contents

1 Introduction

Deliverable 5.7 aims at providing an interim report on the improvements in Machine Translation (MT) related to the semantic linking and resolving activities in WP5 (work package 5). These activities include linguistic processors like Word Sense Disambiguation (WSD), Named-Entity Disambiguation (NED) and Coreference resolution, and Linked Open Data (LOD) resources like WordNet or DBpedia[1] (the LOD version of Wikipedia) for the languages covered in WP5: Basque, Bulgarian, Czech, English, Portuguese and Spanish.[2] The strategy is to explore and experiment with techniques that help to improve Machine Translation performance, and carry over the successful ones to the QTLeap pilot systems.

The main activities refer to the resources and tools integrated in the latest QTLeap machine translation engine (Pilot 2, described in Deliverable D2.8), which following the planning in the DoW, include:

- Replacement of words by interlingual concepts, Experiment 5.4.1, in Section 2,
- Enrichment of word representations, Experiment 5.4.2, in Section 3,
- Creation of specialized lexicons from corpora, Experiment 5.4.3, in Section 4.

Each of the experiments above covers a number of more specific experiments, allowing to test different techniques and variations, as well as idiosyncrasies for each language pair. The most important experiments have been performed on the QTLeap MT platforms, to allow easy deployment of the successful lexical semantic techniques in Pilot 2. The only exception was for en→es: The low results available at the time using TectoMT (see Pilot 1 results in D2.4, also in Tables 23) motivated us to perform some en→es experiments on an alternative platform, Moses, in addition to experiments in TectoMT. The main reasons for using the alternative platform were that improvements over an underperforming system are not scientifically informative, and that it is difficult to set up and evaluate progress of complex experiments using lexical semantic information when the MT platform is still under heavy development. Fortunately, at this point, Pilot 2 for en→es is performing well, and therefore future experiments can all be performed on the QTLeap MT platforms.

In addition to the experiments described in the DoW for this deliverable, we report additional experiments. We report on performance improvements when including coreference information (Section 5).

We also report interim work conducted in the context of WP4 (Section 6), even if this work was not to be reported until the final year of the project. The work investigates the impact of analyzing multiword expressions for machine translation (Task 4.2 in the QTLeap DoW). This encompasses the construction of lists of multiword expressions in the QTLeap languages, and the running of an exploratory experiment by integrating this new resource into the TectoMT system. While this work will not be integrated in Pilot 2, the results described here point the way towards a more sophisticated architecture for multiword analysis, which could be part of Pilot 3.

We also describe MT experiments using lexical semantics performed by DFKI on en→de (Section 7), which is a language not covered in WP5. Let us also mention that

[1] http://dbpedia.org

[2] Although not planned in the DoW, some experiments reported in this deliverable also cover German and Dutch.

we experimented with Dutch, another language which was not part of WP5, as reported in Sections 2 and 4.

Finally, we took the opportunity of this deliverable to report the evaluation of the Language Resources and Tools (LRTs) released in Deliverable D5.6. The LRTs themselves were reported in D5.6, but the report on the evaluation was not planned until the final year of the project and we decided that it was more useful to report the evaluation in this deliverable.

The deliverable is structured as follows. We first describe experiments 5.4.1, 5.4.2 and 5.4.3 in Sections 2, 3 and 4, respectively. Additional experiments on coreference, multiwords and the German MT system are reported in Sections 5, 6 and 7. The performance gains for successful experiments which were included in Pilot 2 are reported in Section 8. Section 9 reports the evaluation of the Basque, Czech and Portuguese LRTs. Finally, Section 10 contains the final remarks.

2 Experiments 5.4.1: Replacing words by concepts

This section presents Experiment 5.4.1, where words in the translation models are replaced by interlingual conceptual representations. Given the higher recall and precision of English disambiguation technology (cf. Deliverable 5.4, also Section 9 in this deliverable), we decided to focus on Word Sense Disambiguation (WSD) on the English side. We planned and performed experiments on en→pt and en→nl. Related Section 3 presents Experiment 5.4.2, where concept information is used in addition to words.

Word Sense Disambiguation returns the sense intended in context, where the sense is actually a link to the interlingual concept representation. In QTLeap (as described in Deliverable 5.4), the concepts in the English WordNet [Fellbaum, 1998] act as interlingual concept representations, represented by synset identifiers in the Interlingual Index (ILI). The synset identifiers (synset ID in short) link concepts across WordNets in multiple languages [Vossen, 2004]. They can be also used to refer to Unique Resource Identifiers of the Linked Open Data version of WordNet.[3] In this deliverable, we refer interchangeably to concepts as synset IDs.

Note the relation between words, word senses and concepts: a polysemous word has several word senses, a word sense refers to the concept referred to by the word, and a concept can be lexicalized by several word senses and words. For instance, the word *house* has several senses. The meaning "business firm" is the sense *house#2*, and refers to the concept `wn30-09213565-n`. This concept is lexicalized by *house#2* and *firm#1*. Word senses can be also represented by the concatenation of a word and a concept, e.g *house*`+wn30-09213565-n` and *firm*`+wn30-09213565-n`.

2.1 Using concepts in en→pt TectoMT

Lexical transfer in TectoMT is based on lemma-to-lemma[4] Translation Models (TMs; see deliverables D2.4 and D2.8 for details). However, lemmas are often ambiguous, unlike word senses. In this set of experiments, we therefore use the information from source language WSD in the TectoMT transfer. The experiments have been carried out on en→pt translation, as the results of en→pt TectoMT for Pilot1 were satisfactory.

To obtain English word senses, we made use of the UKB system [Agirre and Soroa, 2009], a collection of tools and algorithms for performing graph-based WSD over a pre-existing knowledge base. For a word in a context, UKB outputs the intended concept, represented as a synset. This identifier was then utilized in the transfer in two alternative ways:

- Replacing source lemmas with synset IDs (e.g. *house* by `wn30-09213565-n`).
- Replacing source lemmas with word senses (e.g. *house* by *house*`+wn30-09213565-n`).

For the purpose of these experiments, the original lemma-to-lemma translation models (TMs), including both the Dictionary TM and Discriminative TM,[5] needed to be replaced with a newly trained model that used source language synset IDs. In order to speed up the experimentation cycle, all the TMs used in these experiments were trained on a small

[3] http://linghub.lider-project.eu/datahub/vu-wordnet

[4] In fact, these models contain tectogrammatical lemmas (t-lemmas). As it rarely differs from morphological lemmas, we use the term *lemma* in this document, instead.

[5] See D2.4 for more information of the TectoMT architecture.

in-domain corpus, comprising about 16,000 paired entries – 2000 paired sentences from the QTLeap corpus (Batch 1 questions and answers) and about 14,000 paired terms from the localized terminology data for Microsoft[6] (13,000 entries) and LibreOffice[7] (995 entries).

The results of the two approaches measured by BLEU on the Batch2a dataset are shown in Table 1. The system with lemma-to-lemma TMs served as a baseline.

Method	BLEU
Baseline	21.67
Replacing with synset IDs	20.46
Replacing with word senses	19.86

Table 1: The BLEU scores (Batch2a) of en→pt translation using concepts instead of source lemmas in translation models.

A possible explanation for the disappointing results using these approaches is that merely substituting the lemma does not incorporate information which is useful for choosing the appropriate translation. Discriminative TMs have the power to include more powerful contextual features, which we test in Section 3.

2.2 Using concepts in en→nl TectoMT

In coordination with the en→pt experiment, a similar set of experiments was performed for the translation from English to Dutch. Synset IDs returned by UKB were used in the same two simple ways:

- Replacing source lemmas with synset IDs
- Replacing source lemmas with word senses

We used the same training data as we used for the en→nl Pilot 0 experiments (Dutch parallel corpus [Macken et al., 2007, DPC], and KDE localizations [Tiedemann, 2009]). As a baseline, we used the standard version of the TectoMT pipeline with lemma-to-lemma TMs.[8]

The results of the two approaches measured by BLEU on the Batch2a dataset are shown in Table 2.

Method	BLEU
Baseline	23.88
Replacing with synset IDs	19.63
Replacing with word senses	21.67

Table 2: The BLEU scores (Batch2a) of the en→nltranslation using concepts instead of source lemmas in translation models.

As can be concluded from these results, similarly to en→pt, the somewhat naive approach to replace concepts by lemmas does not work at all in en→nl. Using word

[6] Available from: http://www.microsoft.com/Language/en-US/Terminology.aspx

[7] Available from: https://www.libreoffice.org/community/localization

[8] Note that it performs somewhat better than the system we used for Pilot 1, due to a variety of minor improvements, and because we now use the DPC and KDE corpora for training the translation models.

senses instead of lemmas gives somewhat better results but is also clearly worse than the baseline. There are several reasons for this. Both Treex as well as the Alpino generator that is part of the pipeline include rules and heuristics that refer to particular lemmas. Such rules and heuristics no longer work and would need to be adapted to take into account the different set-up. A more fundamental reason for the disappointing results is related to the very specific IT domain of the QTLeap corpus, in combination with the observation that we already use in-domain training data for the translation models (KDE).

2.3 Summary of Experiment 5.4.1

The results gathered in the experiments on two language pairs indicate that naively substituting lemmas with their word senses does not improve results in any of the two pairs. A possible explanation for the disappointing results using these approaches is that merely substituting the lemma does not incorporate information which is useful for choosing the appropriate translation. Discriminative TMs have the power to include more powerful contextual features, which we test in Section 3.

Another explanation might stem from the fact that several TectoMT modules include rules and heuristics that refer to particular lemmas. Such rules and heuristics no longer work and would need to be adapted to take into account the different set-up. Instead of replacing lemmas with concepts, keeping the lemmas and enriching the word representation with concept information would fix this problem, as shown in the next Section.

Finally, the very specific IT domain of the QTLeap corpus, in combination with the observation that we already use in-domain training data for the translation models, could be an additional complicating factor.

3 Experiments 5.4.2: Enriching word representations

The goal of Experiment 5.4.2 was to improve upon the experiment 5.4.1 by enriching word (and lemma) representations with concept information, as well as probability vectors returned by Named-Entity Recognition and Classification (NERC), Named-Entity Disambiguation (NED) and Word Sense Disambiguation (WSD) software. This section will focus mainly on WSD, and Section 4.4 will report work with named entities. The relatively negative results using NERC made us postpone NED experiments for the time being.

We set up several experiments to tackle some of the issues recognized in Experiment 5.4.1. The main idea is to use conceptual information to enrich, not substitute, lemmas. Knowing the word sense of a source word can be useful for translating, but conceptual information can be also useful as contextual a feature. In other words, knowing the semantics of surrounding words can help to translate a source word.

We designed several alternative methods to test this hypothesis, and applied them to different MT platforms and languages, as follows:

- Explore the contribution of word senses returned by UKB on the Discriminative Translation Model (TM) available in TectoMT, which can be used to incorporate conceptual information of both source word and surrounding words as contextual features. We checked different options of word sense representations and a domain-adapted version of UKB. This strategy was followed for en→pt in the experiment described in Section 3.1.

- Explore the contribution of words senses returned by UKB on the Deep Factored MT platform for en↔bg. Instead of using word senses, we experimented with ad-hoc grouping of synsets that share the same target lemma, as this would improve the impact in the translation model. In addition an alternative method for domain-adapted UKB was tested. The results are available in both directions, as reported in Section 3.2.

- Explore the contribution of an alternative WSD system on Moses, where the factors available in Moses are used to represent each token both as a word and as a concept. This experiment was performed for en→es, and described in Section 3.3.

The en→pt and en→es experiments attained some success, as we will see, but their use of WSD is based on a winner-takes-all strategy. In this strategy a single concept is selected for each word, and error can be propagated. We thus designed two additional experiments:

- The first experiment seeks to go beyond the winner-takes-all strategy, where we model the probability vectors returned by the WSD component for the concepts instead. Preliminary experiments using full probability vectors from WSD showed that the current machinery of TectoMT, which needs to build a separate Discriminative TM for each source lemma, does not yield further improvements. We concluded that new transduction mechanisms are needed. For instance, using a single Discriminative TM for all words might be the key to better profit from the additional information. This experiment will be pursued further in the next set of experiments in WP5.

- The second experiment explores the use of distributional semantic vectors (also known as word embeddings), which enrich word representations with a vector of weights which captures the distributional behavior of words [Mikolov et al., 2013b]. This experiment is described in Section 3.4.

3.1 Enriching words with concepts in en→pt TectoMT

As we found no improvement by replacement (in Section 2.1), we tried including synset IDs as additional contextual features in the lemma-to-lemma Discriminative TM. Furthermore, we also introduced corresponding WordNet supersense identifiers. Supersenses are related to the 45 semantic files by which synset identifiers are organized in WordNet, allowing to generalize across semantic classes like PEOPLE, GROUP or ARTIFACT. For instance, the "business firm" sense of word *house* refers to a concept belonging to the GROUP class. We can thus see the GROUP semantic class as the set of all concepts that belong to that class. Supersenses can also be used as coarse-grained senses, e.g. the supersense `house+GROUP` refers to all seven senses of *house* which belong to the GROUP class. If house has 12 senses in WordNet at the synset ID level (fine-grained), it only has 5 senses at the supersense level (coarse-grained). In this deliverable, we refer with supersense ID to the semantic class (e.g. GROUP).

The contextual features were extracted for the English word to be translated (represented by a node in a tectogrammatical tree), but also for its syntactic parent, and its direct left and right sibling, according to the syntactic analysis tree. We experimented with various combinations of these features. For each combination, a Discriminative TM had to be trained for each target lemma. Table 3 shows the BLEU scores achieved with these additional features, with negligible improvements.

We began our experimentation by training over a small, in-domain corpus consisting of the 2000 questions and answers from Batch1, supported by a number of aligned terms sourced from the localized terminology data of Microsoft (13,000 terms) and LibreOffice (995 terms). The resulting in-domain corpus thus comprises approximately 16,000 paired segments, of which 2000 are full sentences and approximately 14,000 are paired terms.

	Node	+Parent	+Siblings	All
Baseline		21.67		
Synset IDs	21.69	21.61	21.68	21.62
Supersense IDs	21.64	21.60	21.62	21.58
Both	21.61	21.61	21.63	21.53

Table 3: The BLEU scores (Batch2a) of the en→pt translation using WordNet information as features in the lemma-to-lemma Discriminative TM.

Domain adaptation to WSD. One of the causes of the poor performance could be that the WSD algorithm does not expect domain-specific text, as pointed out in Section 2.3. We thus designed a domain-specific adaptation of the WSD process. First, a collection of nearest neighbors for terms of interest from the training corpus was created from a pre-existing thesaurus of technological terms (provided by UPV/EHU). The domain-specific thesaurus was extracted by collecting vector representations of words from an automatically-built corpus of 109M words, comprising 209,000 information technology articles and documents extracted from Wikipedia, plus KDE and OpenOffice manuals. The

final thesaurus of each target word is comprised of the 50 most similar words, according to cosine similarity between the vector representations. Given a word[9] from the input sentence, its 20 most similar words at most are retrieved from the collection and included as extra entries in the input context file used by UKB for WSD.

Using this domain-specific adaptation of UKB to perform the WSD process, and repeating the previous approach of adding synset and supersense IDs of source language nodes as features to the lemma-to-lemma Discriminative TM, the system performs in terms of BLEU as shown in Table 4.

Method	Node	+Parent	+Siblings	All
Baseline		21.67		
Synset IDs	21.68	21.63	21.71	21.65
Supersense IDs	21.68	21.64	21.60	21.64
Both	21.67	21.58	21.62	21.54

Table 4: The BLEU scores (Batch2a) of the en→pt translation using WordNet information as features in the lemma-to-lemma Discriminative TM with a domain-adapted WSD.

These results suggest some potential promise in performing the proposed WSD domain adaptation, and then using synset IDs of the current node and its siblings as additional features to the Discriminative TM. This setting has exhibited a small improvement over the baseline.

Training on large open-domain data. We expected that the positive influence of these features on results might be bigger when training the MT engine on a larger, open-domain corpus.[10] Thus, we also ran the experiments using TMs trained on Europarl, containing 1.9 million English-Portuguese sentence pairs.

Method	Node	+Parent	+Siblings	All
Baseline		18.31		
Synset IDs	18.43*	18.45*	18.46*	18.35
Supersense IDs	18.44*	18.30	18.44*	18.46*
Both	18.34	18.50*	18.41*	18.37

Table 5: The BLEU scores (Batch2a) of the en→pt translation using TMs trained in the same way as in Table 4, but on the Europarl data. The symbol * denotes statistically significant ($p < 0.05$) improvement compared to the baseline.

Results in Table 5 show that several configurations outperform the baseline, where most of them are significantly better ($p < 0.05$). One can also notice that using synset IDs as features consistently improves the baseline, regardless of which surrounding nodes they were extracted from. The score is also consistently improved if a feature for the current node and its siblings is added, no matter if it is a synset or supersense ID feature.

[9] We allowed only single-word terms in the current implementation.

[10] We still test on the IT-domain Batch2a, so the BLEU results in Table 5 (training on bigger out-of-domain data) are lower than in Table 4 (training on smaller in-domain data). In the final Pilot 2, we use TM interpolation to take advantage of both training data sets (cf. Section 4.3.)

Comparing the output of the baseline Discriminative TM against the highest scoring Discriminative TM with WSD as features (both synset and supersense IDs from the current node and its parent as features), there are a number of examples where the lexical choice in the WSD model has been improved.

Given a phrase such as "*click the right mouse button*" (with reference translation "*clique com o botão direito do rato*"), the baseline model outputs "*clique no correcto botão de rato*" while the WSD model outputs "*clique no direito botão de rato*". In the baseline model, the word *right* has been translated as *correto* (*right* in the sense of being correct), while the WSD model has made the better lexical choice of *direito* (*right* in the sense of being the opposite of left).

Another example is the phrase "*allows storage and file creation*" (with reference translation "*permite o armazenamento e criação de ficheiros*"), for which the baseline model outputs "*permite armazenamento e criação de processo*" while the WSD model outputs "*permite armazenamento e criação de ficheiro*". In the baseline model, the word *file* has been translated as *processo* (*file* in the sense of a process), while the WSD model has made the better lexical choice of *ficheiro* (the Portuguese word typically associated with computer files).

Of course, there are also examples of less optimal changes in lexical choice, even from the better performing models. For example, considering the phrase "*you will need to go to the menu Insert> Picture*" (with reference translation "*terá de ir ao menu Inserir > Imagem*"), the baseline model outputs "*terá de ir à menu inserção > imagem*", while the highest scoring model with WSD as features model outputs "*terá de deslocar à menu inserir > imagem*". While the WSD model has produced one improved lexical choice in the case of *inserir* as opposed to *inserção*, it has also made a less optimal choice in selecting *deslocar* instead of *ir*. This example highlights the delicate interplay between the different types of word sense information and the node types to which it was added, and their subsequent effects on lexical choice.

Consequently, we eventually chose the following feature sets for the final Pilot 2 experiments:

- Add synset IDs from the current node and its siblings as a feature – synset(node,sibling)
- Add both synset and supersense IDs from the current node and its parent as a feature – synset&supersense(node,parent)

While the latter feature set achieves the highest score, the former is a meeting point of the most what we considered to be the most stable row and the most stable column (there is less variation in results going across the 'synset' row than the 'supersense' or 'both' rows, and less variation going down the 'node,siblings' column than with the other columns). We also noted that synset(node,sibling) scored well throughout our experimentation. These experiments are reported in Table 23 in Section 8.

3.2 Enriching words with concepts in en↔bg Deep Factored MT

The experiments reported in this section exploit the factoring capabilities of the Deep Factored MT platform for en↔bg, which is based on Moses to enrich words with concept information. The semantic information is taken from the Bulgarian and English wordnets[11] in the form of interlingual synset IDs. We used the concept information returned

[11]Bulgarian WordNet was constructed within the project and aligned to the English WordNet.

by the WSD software for the source text in two ways: use synset ID directly as a factor or choose a *representative lemma* in the target language for the synset and use this representative lemma as a factor (in addition to the source word-form factor).

The motivation for using representative lemma in the target language is the hope to unify the various synset IDs with similar translation in the target language. For example, in the en→bg direction, the two concepts referred by *donor*: `wn30-10025730-n` ("person who makes a gift of property") and `wn30-10026058-n` ("a medical term denoting someone who gives blood or tissue or an organ to be used in another person") are very close each to other, but they have the same translation in Bulgarian in both corresponding synsets: ***донор***. The representative word is selected on the basis of a frequency list of Bulgarian lemmas constructed over large corpora (70 million words).

We performed three experiments: using synset IDs returned by the WSD software (ExpA); using representative target language lemmas for the synsets returned by the WSD software (ExpB); using representative target lemmas where the WSD software is run on domain-adapted wordnets, which have been extended with domain gazetteers and terms (ExpC).

The experiments for en→bg have been performed through the following steps: (1) annotation of the English text with the IXA[12] pipeline including tokenization, sentence splitting, part-of-speech tagging and word sense annotation using UKB; (2) substitution of the English word form with the synset (in ExpA) or Bulgarian representative lemma (in ExpB and ExpC); and (3) factored model in the Moses system. In direction bg→en we performed similar processing, but using the Bulgarian analysis pipeline prepared in QTLeap. Additionally, we provide part-of-speech tags[13] (PoS) from the pipeline as well as the source-language lemma as factors for Moses.[14] The PoS factor is important because Bulgarian is morphologically rich.

As an example, the procedure we performed with respect to the training, testing and tuning of the Moses system is as follows:

English sentence:
This is real progress ...

English sentence with factors:
this|this|dt is|be|vbz реален|real|jj напредък|progress|nn .|.|.

Bulgarian sentence with factors:
това|това|pd е|съм|vx реален|реален|a напредък|напредък|nc .|.|pu

Bulgarian sentence:
Това е реален напредък.

In order to adapt the semantic processing, we incorporated a Linked Open Data resource DBPedia in the en↔bg experiments via a mapping of the DBpedia ontology to WordNet. Our goal was to use again IXA pipeline for the WSD task similarly to lexical semantics experiments. Unfortunately, the DBpedia ontology contains very few relevant classes like, for instance *software, website, database*. For that reason, we decided to use

[12] http://ixa2.si.ehu.es/ixa-pipes/

[13] In our case PoS tags include some morphosyntactic features.

[14] So although the source-language word form is substituted (if a synset for it is detected), its lexical information is still present in the lemma, which is used as an additional factor. Thus all three ExpA, ExpB and ExpC fit to Experiment 5.4.2, where words are enriched with additional features (factors in case of Moses), rather than to Experiment 5.4.1, where words were naively substituted with word senses.

an additional ontology created in a previous European project, LT4eL,[15] which covers about 1500 domain classes in the domain of Information technology. This ontology is already aligned to OntoWordNet [Gangemi et al., 2003], which is a basis for extension of the existing wordnets as used by the WSD UKB system.

In ExpC, in order to further adapt the wordnets to the domain by adding new concepts, instances and relations, we annotated manually the test texts from Batch 1.

The Batch 1 dataset represents domain specific texts in the IT domain and consists of questions and their answers. It contains 29,901 tokens, which build 2,579 sentences. Some of the questions and answers consist of more than one sentence. The tokenization was performed automatically, and then post-edited manually. The texts were morphologically annotated in an automatic way, but then checked manually. There were 7,788 morphological ambiguities, which were also disambiguated manually. Thus, this annotation is considered a gold standard.

For the morphological annotation an extensive dictionary of Bulgarian Inflectional Morphology, containing about 100,000 lexemes, was used, but some of the lexemes in Batch1 were not recognized by the dictionary. These are domain specific words like ***драйвер*** (*driver*), ***плъгин*** (*plugin*). After the annotation, more than 200 lexemes were added into the dictionary.

Also, domain specific named entities have been added to the setting, such as Windows, Excel, etc. These names are classified with respect to their concept. For example, "Excel Options" is classified as an icon on the user interface, "Sent Items" is classified as a folder, etc. Other concepts used for classification of the Domain Named Entities are "command", "product", "company", "keyboard button", etc.

All the 2,579 sentences were then annotated with senses. The sense annotation includes the following parts-of-speech: verbs, nouns, adjectives and adverbs. The senses were taken from the BulTreeBank WordNet.[16] This annotation was manually checked. Additionally to the senses from WordNet, we used a domain specific dictionary with more than 900 words in the IT domain. Most of the senses in the domain specific dictionary represent new senses to ambiguous words like ***пиша*** (*type*) with a definition "Набирам текст" ("write by means of a keyboard with types").

One third of the words in the corpus happen to have domain specific senses. Among the words with domain specific senses, 209 are mapped to Wikipedia and have URIs. But there are also words with domain specific senses that do not have a representation in Wikipedia, most of them are verbs like ***сканирам*** (*scan*) and ***деинсталирам*** (*uninstall*). 651 words from the domain specific dictionary were added to the BulTreeBank WordNet and then mapped to Princeton WordNet. In this way, we have extended the UKB knowledge graph with domain nodes.

From the annotated corpus, we have extracted co-occurrence relations between newly added nodes. In some cases, however, the extracted relations are between the new nodes and the nodes in the existing knowledge graph.

After we performed the substitution of synsets with selected representative target-language lemmas, we trained the Deep Factor MT platform with the following factors: SubstitutedWordform[17], Lemma, PoS tag.[18] As a baseline we use the results from the

[15] http://www.lt4el.eu/

[16] http://compling.hss.ntu.edu.sg/omw/

[17] For some word forms like prepositions, conjunctions, etc. the original word form is kept.

[18] The parameters for training the Moses system are: `--translation-factors 0,2-0,2+1,2-0,2 --decoding-steps t0:t1`

evaluation of Pilot 0 model on Batch3a (en→bg direction) and Batch3q (bg→en direction). Table 6 presents the results.

System	Source factors	Domain terms	en→bg BLEU	en→bg NIST	bg→en BLEU	bg→en NIST
baseline	form	no	17.72	–	22.56	–
ExpA	synset|form, lemma, PoS	no	16.23	4.81	16.72	4.99
ExpB	repres-lemma|form, lemma, PoS	no	17.23	4.95	20.05	5.61
ExpC	repres-lemma|form, lemma, PoS	yes	17.41	4.98	19.92	5.58

Table 6: BLEU and NIST results of en↔bg experiments with concepts on Batch3 (Batch3a for en→bg and Batch3q for bg→en). The baseline is Pilot 0, where no synsets are used. In ExpA, the synset ID is added (if it exists). In ExpB, *repres-lemma* is added, which is a representative target-language lemma for the given (source-language) synset. ExpC is same as ExpB, but the WSD used was enriched by domain terms.

All three en↔bg experiments with lexical semantics (ExpA, ExpB and ExpC) show a drop in the results with respect to the baseline.

The lack of improvement for the en→bg direction (QTLeap answers) contrasts with the improvement for en→pt reported in the previous section. The use of a discriminative classifier instead of a factored model could be the cause, but also the method to encode word senses.

The lack of improvement for bg→en for queries contrasts with the positive results on queries for en→es (Section 3.3) and en→de (Section 7). The cause might be due to the different technique used to encode word sense information. Another alternative explanation could be that the current quality of the Bulgarian resources and WSD module is not satisfactory.

3.3 Enriching words with concepts in en→es Moses

The experiments described here enrich word representations as in the previous section, but with some differences, as follows:

- We test a different WSD software (SuperSense Tagger instead of UKB)
- We use a statistical machine translation engine (Moses instead of TectoMT)
- We use factors in Moses to represent words and concepts separately

In order to test alternative WSD software, we have run a SuperSense Tagger on the English part of the train data. SuperSense Tagger is a sequential labeller that deploys a discriminatively trained Hidden Markov Model. The model can be seen as a perceptron-trained Hidden Markov Model [Collins, 2002] that jointly models observations and label sequences. The used SuperSense tagger is a reimplementation of Ciaramita and Altun [2006] provided by Michael Heilman.[19] Following common practice with supersense tagging, the tagger learns 83 labels: 41 supersense IDs,[20] with (B) beginning and (I) continuation as prefixes, plus no label category (O). The features used in the implementation are common in WSD and NERC.

[19] http://www.ark.cs.cmu.edu/mheilman/questions/SupersenseTagger-10-01-12.tar.gz

[20] This is the full set of 45 classes, discarding the label for adverbs and the three classes for adjectives

The supersense IDs have been added as factors, and several experiments have been performed using different factor paths. In the best configuration, two alternative translation paths are combined (see Figure 1): a direct translation path (where target words are generated from source words) and a sense-augmented translation path, where in addition to source words supersense IDs are also taken into account. Translations generated by both translation paths are combined at the decoding phase to generate the final translation. [21] The figure is akin to the following: Word ← word, Sense ← word (alt path) (cf. Section 7).

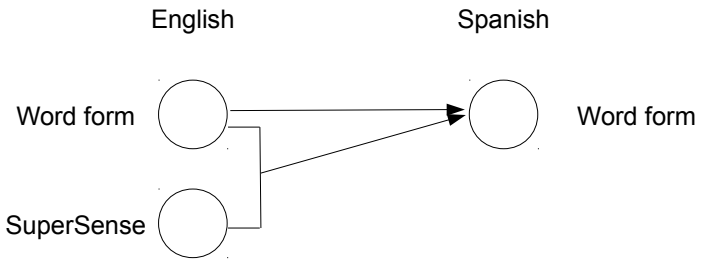


Figure 1: Best factor configuration using supersenses

This supersense-augmented Moses system was compared with the non-factored Moses on three test sets: Batch2q, and Batch2a, both representing the IT-domain, and the test set from WMT 2013 Shared Task on machine translation,[22] representing the news domain. The BLEU scores of the en→es translation are shown in Table 7, where we can see small improvements in two out of the three datasets, queries and News.

Method	Batch2a	Batch2q	WMT13
Moses baseline	33.36	39.47	26.04
Supersense as factor	33.25	39.74	26.38

Table 7: The BLEU scores on en→es when using Supersense Tagger supersense IDs as factors of Moses.

Aware of the limitations of the factored approach, we are currently doing experiments on the use of purpose-built classifiers. These classifiers return, for each entry in the phrase table, context-based probabilities for possible translations. This approach has been successfully used in a similar setting using word sense induction techniques for Chinese to English translation [Xiong and Zhang, 2014]. In our case, we use supersense IDs instead of induced word senses, and a powerful all-words classifier[23] instead of one classifier per source phrase. In addition to the best concept, we have also represented each token with features which capture the probability vector returned by the WSD system. More specifically, each token is enriched with a vector of 41 weights (one per supersense ID), where each weight corresponds to the probability assigned by the Supersense tagger for

[21] These are the respective parameters: `-translation-factors 0-0+0,1-0 -decoding-steps t0:t1.`
[22] http://www.statmt.org/wmt13/test.tgz
[23] https://github.com/JohnLangford/vowpal_wabbit/

the token to have the supersense ID. These experiments are on-going, due to the time-consuming training process.

3.4 Enriching words with distributional representations in en→cs TectoMT

In addition to full probability vectors, we also started to run preliminary experiments with distributional semantic vectors (also known as word embeddings), which enrich word representations with a vector capturing the distributional behavior of words [Mikolov et al., 2013b]. We encoded the enriched representations as additional features for the Discriminative TM. Based on these preliminary experiments (on small-scale data), we observed the following:

1. We cannot use embeddings of the word being translated because of the limitations of the current Discriminative TM, which trains a separate MaxEnt model for each source lemma. So all the competing translations would have the same embedding features coming from the same source lemma and they would have no effect on the translation. One solution is to train one huge model for translation of all lemmas, which would allow us to use so-called label-dependent features. So the embedding features of the word being translated would be shared over all source lemmas and we could model, how individual dimensions of the embeddings correspond with various translations.

2. With the current Discriminative TM, it is difficult to use embeddings from the syntactic context (dependency parent and children) of the word being translated because this results in many extra features (in addition to the existing features), which is on the border of the technical capabilities of the current Discriminative TM. The actual features used internally in MaxEnt would be of form (source_lemma, target_lemma, context_word[i].embedding[j]),[24] where i is the index of the word (e.g. parent, first child, second child,...) and j is the j-th dimension of the embedding vector (we have used 300 dimensions in our experiments).

3. Ideally, we would like to use also embeddings from the target lemmas and include them into the (label-dependent) features. We would also like to model interactions between different dimensions of the word embeddings. A naive approach would add too many too sparse features,[25] so we plan to use neural networks, so we can model the non-liner interactions between the features, but we don't need to model the full Cartesian product.

[24] The source_lemma is now present only implicitly, as a separate model is currently trained for each source lemma.

[25] Suppose our embeddings have 300 dimensions and we limit the number of target-language lemmas to 10,000. For modeling interactions between pairs of source-embedding dimension and the target lemmas, we need $10,000 \times \frac{300 \times 300}{2} = 450$ million features. If we want to add also embeddings from the parent and children (not distinguishing which children the embeddings came from), we would need 300×300 times more features (40,500 billion features). When using pairs of target-word embedding dimensions (90,000) instead of target lemmas (10,000), we would need 9 times more features. So unlike in point 2, this is beyond the technical capabilities of any toolkit we know, not only the current Discriminative TM. It is also problematic from the machine-learning point of view because we do not have enough training examples to train a model with so many features, so overfitting would be a problem even with regularization used.

4. Even without increasing the number of features, training of the current MaxEnt models is too slow (e.g. training on CzEng [Bojar et al., 2012], a parallel treebank with over 200 million words on both sides, takes about 4 days on a cluster with 200 machines).

As a solution to all the four problems, we plan to use Vowpal Wabbit[26] machine learning toolkit for Discriminative TMs instead of the current MaxEnt models. Vowpal Wabbit enables also using a simple neural network with one hidden layer. In our preliminary experiments, we were able to replicate (and even outperform) en→cs results on the News domain. The training (after extracting features) took 2 hours on a single two-core machine. However, integrating the word embedding features into Vowpal Wabbit in an efficient way still remains to be done for Pilot 3.

3.5 Summary of Experiment 5.4.2

Including sense information as features of the Discriminative TM of TectoMT was tested on en→pt, and yields positive gains when applied on Europarl data using a domain-adapted variant of the WSD algorithm. The use of word sense information of context words (adjacent words in the syntactic tree) seems to indicate that disambiguating the words in the context of the source lemma is more useful than using the word sense information of the source lemma directly. The current Discriminative TM is somewhat limited, as it needs to train a separate classifier for each target lemma, and is thus unable to generalize across lemmas which are closely related. For instance, in order to translate an occurrence of the verb *run* in the IT domain (translated to Spanish as *correr* in the physical sense but as *ejecutar* in the IT sense), a classifier which has very few examples of *run* in the IT sense would need a strong signal to realize that the test context is related to IT. We hypothesize that a single Discriminative TM which takes into account all contexts of all target words has better chances to successfully use word sense and related semantic class information, compared to building separate models for each source lemma.

The experiments for en↔bg with ad hoc grouping of synsets that share the same lemmas like "donor" mentioned above did not prove a good way to incorporate lexical semantics in factored MT. One possible explanation of this is that the performance of the sense annotation module is not high enough. Especially for Bulgarian it is around 68 % on a gold standard corpus. The addition of domain terms improves a little the result, but obviously more work is necessary to show the positive effect of lexical semantics in these MT settings.

Regarding the en→es experiment, the good results using factors in Moses are a good indication that conceptual information can be successfully captured using a Supersense Tagger. This opens the opportunity to combine the output of UKB with that of the Supersense tagger in order to improve WSD results.

Regarding datasets, it seems that WSD is specially helpful for QTLeap queries and the news domain, as shown by the improvements in en→es. This is a promising direction for non-English WSD, as the QTLeap evaluation scenario involves the translation of queries with English as the target language. The lack of improvement for QTLeap answers using factors in en→es contrasts with the improvements in en→pt when using discriminative classifiers, which hints at the superiority of that technique to combine WSD information into the MT engine.

[26] https://github.com/JohnLangford/vowpal_wabbit

Finally, we have started to try new translation MT models, using one single Discriminative TM in order to cope with the richer word representations (probability vectors from WSD and distributional representations of words), which tend to break current lemma-by-lemma Discriminative TM technology. From another perspective, we can argue that the WSD algorithm is enriching the MT model with information which is orthogonal to that available in the parallel corpora. Following this perspective, the semantic information will be specially relevant for those source lemmas which occur infrequently in the training corpus of the MT system. Again, a single Discriminative TM seems more appropriate. This is ongoing work.

4 Experiments 5.4.3: creation of specialized lexicons from corpora

In this Section, we present the experiments related to the creation of specialized lexicons. The error analysis of the Pilot 1 systems revealed that a substantial number of errors in the QTLeap domain are caused by wrong handling of named entities (NEs). Even though in some of the translation directions the tools for NERC (cf. Deliverable 5.4) were used, they are not able to cover specific types of NEs one can encounter in the IT domain of the QTLeap corpus. These include URLs, shell commands, and code snippets on the one hand – all addressed by the HideIT machinery (see Section 4.1), and the special types of text contained in software, e.g. menu items, button names, their sequences, and messages on the other hand – all addressed by specialized lexicons, also called gazetteers.[27] Section 4.2 describes the process to construct such gazetteers from multilingual corpora like localization files and Wikipedia.

Regarding the MT machinery, whereas the former group of entities (URLs,…) must be recognized and usually stays untranslated, the latter group (menu items,…) should be translated according to the localization rules of the software it originates from. Nevertheless, both groups consist of expressions that are rarely inflected even in inflectional languages, such as Czech. That is the reason why both approaches are applied on a tokenized text before any linguistic analysis is performed. When combining gazetteers with translation models, several translations might be competing. Section 4.3 presents an alternative solution how to use in-domain knowledge and adapt the system for IT domain – TM interpolation.

Apart from the entities relevant to the QTLeap domain, we have also tested the contribution of the named entities detected by NERC software, such as people, location and organization names. Their relevance in the QTLeap domain is negligible, but they are important when translating News text. Section 4.4 presents a dedicated module to translate such named entities.

Finally, Section 4.5 presents a method to gather automatically domain-specific parallel corpora from Wikipedia.

4.1 "Fixed" entities (HideIT)

To address the problem of entities that do not require translation, e.g. URLs,[28] shell commands and code snippets, we use a rule-based machinery called HideIT in TectoMT.

The HideIT machinery consists of two blocks. The first one is applied at the very beginning of the translation pipeline: just after tokenization of the source text, before any linguistic processing is applied. The block attempts to recognize "fixed" entities by heuristics manually gathered on Batch 1 corpus. The recognized entities are then removed and replaced by an appropriate placeholder (e.g. `xxxCMDxx` and `xxxURLxxx` for shell command and URL, respectively), while storing the actual values in the metadata.

The second block is applied at the very end of the translation pipeline. Given the placeholders, it extracts the stored values from the metadata and restores the entities

[27] As usual in NERC research, *gazetteer* means a list of named entities of a given type, not only geographic names.

[28] Note that even URLs and e-mail addresses may be required to be localized, e.g. `www.example.com/en` for English-speaking and `www.example.com/cs` for Czech-speaking users. However, regular expressions seem to be a more appropriate solution than lexicons in such cases. We do not address this problem, yet.

which have been hidden from the main part of the translation pipeline.

These blocks have been used for all translation directions using TectoMT as the Pilot 2 system. Tables 22 and 23 in Section 8 show the effect of switching on the HideIT machinery in Pilot 2 systems. HideIT had negligible effect on X→en translations (because queries do not contain so many "fixed" named entities), but consistently improved en→X translations (about 0.5 BLEU point on average).

4.2 Specialized lexicons (gazetteers)

The technique of gazetteers targets the NEs from the IT domain that need to be translated or localized. Furthermore, they are expected to appear in a fixed inflectional form.[29] It concerns software texts, such as menu items, button names, their sequences, and messages. The property of having a fixed form allows us to apply the technique of matching the expressions from a *specialized lexicon (gazetteer)* in the surface source text and replacing them by their equivalents in the target language. A crucial task is also to identify the source expressions with this behavior.

Lexicon collection and format As the majority of this kind of expressions consists of texts appearing in various software, the straightforward way how to obtain a lexicon of such expressions is to extract it from the freely available software localization files. We designed a general extractor that accepts `.po` localization files and outputs a lexicon. The lexicon is formed by two lists containing corresponding expressions in two languages. Each of the two lists consist of two columns:

1. expression identifier
2. expression itself

The identifier column must be the same for both lists belonging to the same lexicon.

Figure 2 shows a toy English-Czech gazetteer. It consists of two lists, one for each of the languages, which are paired by identical expression identifiers.

Translation using specialized lexicons Translation using gazetteers proceeds in multiple steps:

Matching the lexicon items. This is the most complex stage of the whole process. It is performed just after the tokenization, before any linguistic processing is conducted. Lexicon items are matched in the source tokenized text and the matched items, which can possibly span several neighboring tokens, are replaced by a single-word placeholder.

In the initialization stage, the source language part of the lexicon is loaded and structured in a word-based trie to reduce time complexity of the text search. In the current implementation, if an expression appears more than once in the source gazetteer list, only its first occurrence is stored, regardless what its translation is. Therefore, the performance of gazetteer matching machinery depends on the ordering of the gazetteer lists. A trie built from the English list of the toy English-Czech gazetteer is depicted in Figure 3. Note that the `kde_7` item is not represented in the trie, since the slot is already occupied by the `kde_3` item.

[29] If they appear in a different form, only the forms contained in a gazetteer list (base forms, mostly) will be treated this way.

English list: `toy.en-cs.en.gaz`

```
liboff_1    Accessories
liboff_2    Start at
kde_1       Programs
kde_2       System tools
kde_3       Start
kde_4       Disk
kde_5       PC running on low battery

kde_6       System
kde_7       Start
wiki_1      PC
```

Czech list: `toy.en-cs.cs.gaz`

```
liboff_1    Příslušenství
liboff_2    Začít od
kde_1       Programy
kde_2       Systémové nástroje
kde_3       Spustit
kde_4       Disk
kde_5       Počítač je napájen téměř
             vybitou baterií
kde_6       Systém
kde_7       Start
wiki_1      PC
```

Figure 2: A toy English-Czech gazetteer.

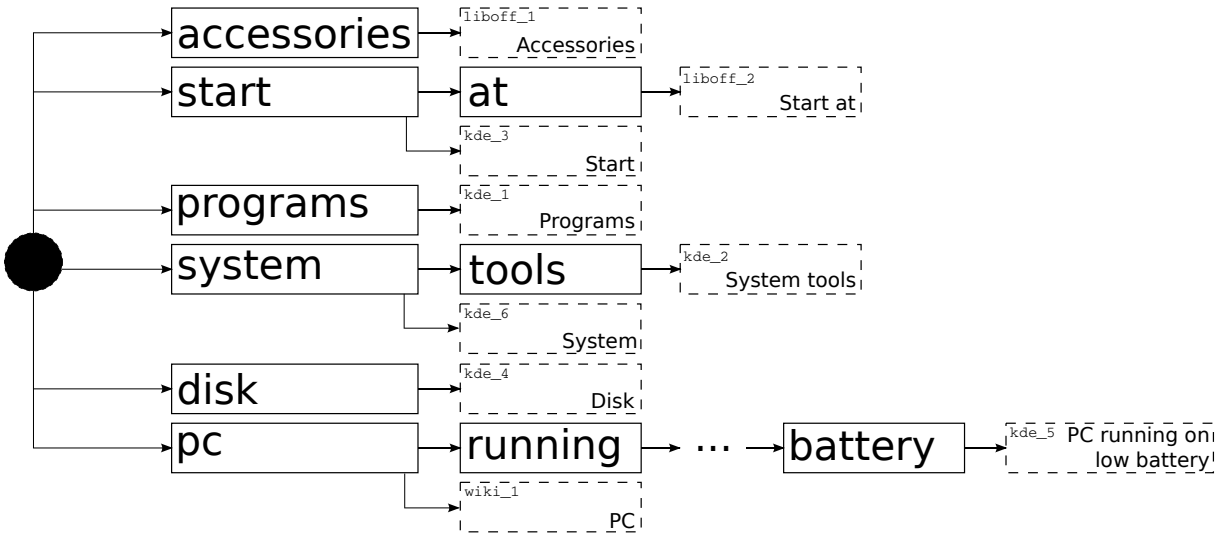


Figure 3: A trie created from the English list of the toy English-Czech gazetteer.

The trie is then used to match the expressions from the list in the source text. The matched expressions might overlap. Thus, every matched expression is assigned a score estimating the extent to which the expression is a named entity. In Figure 4 a sample sentence (a) is shown with the expressions matched and scores assigned (b).

The matches with positive score are ordered by the score and filtered to get non-overlapping matches, taking those with higher score first. The matched words belonging to a single entity are then replaced by a single word. The placeholder word can be specified with respect to the source language, the word *Menu* being the default (see Figure 4c).

As a last step, the neighboring entities are collapsed into one and replaced by the placeholder word. The entities are collapsed also when they are separated by a > symbol (or possibly any other delimiter). This measure is aimed at translation of menu items and button labels sequences, which frequently appear in the QTLeap corpus. After this step, the sample sentence becomes drastically simplified, which should be much easier to process by a part-of-speech tagger and parser (see Figure 4e). However, all the information

necessary to reconstruct the original expressions or their lexicon translations are stored (see Figure 4d).

Translating matched items. The expressions matched in the source language are transferred over the tectogrammatical layer to the target language. Here, the placeholder words are substituted by the expressions from the target language list of the gazetteer, which are looked up using the identifiers coupled with the placeholder words. Possible delimiters are retained. This is performed before any other words are translated. The tectogrammatical representation of the simplified sample English sentence (Figure 4d) is transferred to Czech by translating the gazetteer matches first, followed by lexical choice for the other words and concluded with the synthesis stage (Figure 4g).

a) `To defragment the PC, click Start > Programs > Accessories > System Tools > Disk Defragment.`

b) `To defragment the [PC wiki_1=24], click [Start kde_3=24] > [Programs kde_1=24] > [Accessories liboff_1=24] > [[System kde_6=24] Tools kde_2=44] > [Disk kde_4=24] Defragment.`

c) `To defragment the [Menu wiki_1], click [Menu kde_3] > [Menu kde_1] > [Menu liboff_1] > [Menu kde_2] > [Menu kde_4] Defragment.`

d) `To defragment the [Menu wiki_1], click [Menu kde_3 > kde_1 > liboff_1 > kde_2 > kde_4] Defragment.`

e) `To defragment the Menu, click Menu Defragment.`

f) `To defragment the [PC wiki_1], click [Spustit > Programy > Příslušenství > Systémové nástroje > Disk kde_3 > kde_1 > liboff_1 > kde_2 > kde_4] Defragment.`

g) `Jestli defragmentujete PC, klikněte na Spustit > Programy > Příslušenství > Systémové nástroje > Disk defragmentaci.`

Figure 4: A sample English sentence processed by the English-Czech gazetteer.

Statistics and sources The gazetteers for Czech, Dutch, Basque, Spanish and Portuguese were collected from four different sources: localization files of VLC,[30] LibreOffice,[31] KDE[32] and IT-related Wikipedia articles. In addition, some manual filtering was performed on all the gazetteers, especially the Czech one. The Czech gazetteer was also enriched with some entries frequent in the Batch1 dataset but not covered by any of the sources mentioned above.

For mining IT-related terms from Wikipedia, we adopted the method by Gaudio and Branco [2012]. This method exploits the hierarchical structure of Wikipedia articles. This structure allows for extracting articles on specific topics, selecting the articles directly

[30] http://downloads.videolan.org/pub/videolan/vlc/2.1.5/vlc-2.1.5.tar.xz

[31] http://download.documentfoundation.org/libreoffice/src/4.4.0/libreoffice-translations-4.4.0.3.tar.xz

[32] svn://anonsvn.kde.org/home/kde/branches/stable/l10n-kde4/{cs,nl,es,eu,pt}/messages

linked to a superordinate category. For this purpose, Wikipedia dumps from June 2015 were used for each of the languages, and they were accessed using the Java Wikipedia Library, an open-source, Java-based application programming interface that allows to access all information contained in Wikipedia [Zesch et al., 2008]. Using as starting point the most generic categories in the IT field, all the articles linked to this categories and their children were selected. The title of these article were used as entries in the gazetteers. The inter-language links were used to translate the title in the original languages to English. Similar result could be expected if the method was applied to the Linked Open Data version of Wikipedia, DBPedia,

The figures of collected gazetteer entries for all the sources are presented in Table 8.[33] The gazetteers have been released through Meta-Share.[34]

	en-cs	en-nl	en-eu	en-es	en-pt	en-de
KDE	124,188	98,512	70,298	98,510	98,505	—
LibreOffice	75,662	75,457	70,991	75,482	75,743	—
VLC	3,467	6,213	5,548	6,214	6,215	—
Wikipedia	28,196	39,570	1,505	24,610	20,239	23,011
Batch1	3	—	—	—	—	—
Microsoft Terminology	—	—	—	—	—	22,972
Total	231,516	219,752	148,342	204,816	200,702	45,983

Table 8: Source and number of gazetteer entries in each language.

Using all the sources proved to be the most beneficial setting for most translation pairs. The only exceptions are translations from Spanish to English and from Dutch to English. Whereas es→en performs the best with matching only the Wikipedia entries, the optimal performance of nl→en can be reached without gazetteers. These configuration have been used in Tables 22 and 23 in Section 8 for Pilot 2 systems based on TectoMT. As can be seen there, gazetteers helped to improve the translation quality in all TectoMT-based en→X directions (2 BLEU points on average) and also in cs→en and es→en (about 0.7 BLEU point).

4.3 TM Interpolation

As described in the previous section, when combining gazetteers from various sources, they need to be ordered to decide cases when more gazetteers would match a given phrase. Moreover, when gazetteers match a phrase, the standard TectoMT Discriminative TM are bypassed, and cannot influence translation of this phrase. This is the reason why the gazetteers should include only clear-cut cases and why we use heuristic scoring to filter out cases when a phrase is not used as a named entity.

So there is a number of cases which are not solved by gazetteers, but adaptation for IT domain is still needed. Several translations might be competing, some coming from the IT domain, other coming from the general domain.

Instead of using hard-coded rules, TectoMT allows to use TM interpolation. Domain adaptation using interpolation of translation models (general domain and IT domain) is

[33] For the sake of completeness, the table also contains gazetteers for German, which have been used in an approach different from TectoMT, described in Section 7.

[34] http://metashare.metanet4u.eu/go2/qtleap-specialized-lexicons

described in detail in D2.8, Section 2.2.5 because it is closely related to the description of the transfer phase. Interpolation is applied not only on lexical transfer (as gazetteers and HideIT), but also on transfer of formemes. In fact, we observed that IT domain has a different distribution of formeme translation probabilities, so this interpolation is helpful.

In this deliverable (D5.7), we present the results with and without interpolation. Tables 22 and 23 show very good results, with improvement in all translation pairs (1.2 and 1.7 BLEU points on average for X→en and en→X, respectively), most notably en→es (5.1 BLEU points).

4.4 Experiments with Named Entities in en→es

The experiments described here are related to the treatment of domain-specific named entities (NEs) with heuristics and gazetteers as described in the previous sections. The main differences are the following:

- We focus on the news domain, where most of the NEs are of the standard PERSON, LOCATION and ORGANIZATION types, in contrast to the IT domain of the QTLeap corpus, where the relevance of domain-specific entities like URLs, shell commands or menu items is more prominent.
- We use a statistical machine translation engine (Moses instead of TectoMT).

The exploration of this alternative techniques is relevant, given the interest in both the IT and the news domain for QTLeap.

These experiments focus on improving NE translation, a field called name-aware SMT. The most basic approach is to add a devoted NE translation gazetteer to the training data. Pal et al. [2010] report good results using this method. Another common solution is to replace NEs by special tags and translate them in a post-edition step. This approach is similar to the one used in the previous Sections 4.1 and 4.2, but they keep all relevant information in the transfer phase. Okuma et al. [2008] propose replacing source names by high frequency names before applying SMT. In a more sophisticated setting, Li et al. [2013] use a Hierarchical SMT system (HSMT) to integrate a specialized NE translation system, showing relevant improvements in overall translation quality and, particularly, in NE translation when translating from Chinese to English. In this experiment, we replicate their system and analyze how NEs are translated when translating from English to Spanish on the news domain.

Analysis of NE translation in the news domain In order to better understand how traditional SMT systems perform when translating NE from English to Spanish, we carried out a manual analysis over 525 sentences that were randomly taken from the news-test2011 test set as given in WMT 2011 Shared task on machine translation,[35] which we used as our development set. We note that, in some cases, both Spanish and English text seemed to be actual translations from a third language.

We first run the ixa-pipe-nerc NERC system [Agerri et al., 2014] on these sentences, and manually assessed the correctness of each of the 536 NEs that it recognized, as shown in Table 9. We then identified how each of the correctly recognized NEs was translated in the reference translations. We discovered that 1.61% of them were missing in the translations, 3.63% were not translated correctly, and another 2.82% had a meaningful

[35] http://www.statmt.org/wmt11/test.tgz

but indirect translation (e.g. a country name translated as a demonym). This means that, even in the human translation, only 91.94% of the NEs had a correct NE translation in the reference translation.

Correct				Wrong
Person	**Location**	**Organization**	**Misc.**	
123 (22.95%)	184 (34.33%)	132 (24.63%)	57 (10.63%)	40 (7.46%)

Table 9: Distribution of NEs in the development set

We then checked the performance of a baseline system HSMT system trained on Europarl v7 using Moses [Koehn et al., 2007]. Table 10 shows the amount of correctly translated NEs for this system, according to their class and number of occurrences in the training corpus. The results suggest that our baseline system performs relatively well for this task (86% overall), and that the errors are concentrated on NEs with zero or one occurrences (approx. 77% accuracy), with very good performance for NEs occurring more than once. We analyzed the errors and found that 28.17% of them corresponded to untranslated NEs, whereas another 23.94% were caused by proper nouns that were translated as common nouns even though they should have been kept unchanged.

Occurrences	**PER(%)**	**LOC(%)**	**ORG(%)**	**MISC(%)**	**Total(%)**
0	88.46	78.43	63.38	50.00	77.12
1	100.00	85.71	62.50	100.00	77.78
>1	100.00	99.21	92.45	82.61	94.63
Total	90.24	92.93	75.00	77.19	85.69

Table 10: NE translation accuracy in the development set for the baseline HSMT system, split by number of occurrences (rows) and NE type (columns)

In conclusion, we can say that, compared to Chinese-English [Li et al., 2013], the room of improvement is smaller (roughly 15% vs. 30%). We thus decided to focus on OOV (i.e., 0 occurrences in the training data) and hapax legomena (i.e., 1 occurrence in the training data) NEs.

NE-enhanced HSMT system. Our approach for improving NE translation in SMT is based on the framework proposed by Li et al. [2013]. We train a HSMT system with Moses, adapting the training phase to treat each NE class as a non-terminal. Given our analysis (see above), NE occurring more than once are left for the HSMT to handle. In the case of NEs with zero or one occurrences, we use a specialized module to generate additional translations that are added to the phrase table on the fly. This module merges the results of several independent techniques to translate NEs: an automatically extracted dictionary, a human dictionary, Wikipedia-related Linked Open Data resources, leaving the NE unchanged, a special treatment for title + person structures, a rule-based machine translation engine and an SMT system specialized on NE. Each translation technique is given an independent weight, and the system is tuned to optimize these weights.

We used news-test2012[36] as our test set and took 525 random sentences to measure NE translation accuracy and the full test set to calculate the BLEU score. Table 11

[36] http://www.statmt.org/wmt12/test.tgz

shows the results obtained by this system in comparison with the baseline system. Our results show a small but statistically significant improvement of 0.2 BLEU points, but no improvement in terms of NE translation accuracy. Note that 7.17% of the NEs were translated differently. We are currently studying the reasons of the improvement in BLEU.

Method	BLEU	NE translation accuracy
Baseline HSMT	31.01	414 (87.34%)
NE enhanced HSMT	31.21	415 (87.55%)

Table 11: NE translation accuracy and BLEU score in the test set

More information on this work can be found in Artetxe et al. [2015].

4.5 Experiments on domain corpora from Wikipedia

The experiments described here are related to the construction of gazetteers described in 4.2. The main differences are the following:

- We focus on producing parallel corpora instead of gazetteers, exploiting article content in Wikipedia.
- We use a statistical machine translation engine (Moses instead of TectoMT).

This work is related to domain adaptation, which has recently gained interest in statistical machine translation to cope with the performance drop observed when testing conditions deviate from training conditions. The basic idea is that in-domain training data can be exploited to adapt all components of an already developed system. Previous work showed small performance gains by adapting from limited in-domain bilingual data [Bertoldi and Federico, 2009]. Some of the works improved MT using Wikipedia [e.g., Gupta et al., 2013], but previous work does not identify domain-related parallel corpora from Wikipedia, and thus only use small units, mainly named entities.

For the English–Spanish language pair, we extracted domain specific parallel corpora from Wikipedia and used it as additional corpora to train an SMT system. We make use of the natural function of Wikipedia as a source of multilingual data to gather corpora for multiple domains. Our methodology is divided in three steps: (i) selecting those Wikipedia articles related to the specific domains for every language independently; (ii) extracting comparable corpora in these domains using Wikipedia's language links; and (iii) identify parallel sentences from those comparable corpora

Wikipedia as a source for domain corpora. In order to select articles from a given domain, we take advantage of the graph structure in Wikipedia. the method could have been applied on DBPedia, the Linked Open Data relative of Wikipedia, and the result would have been the same.

Wikipedia's users can categorize the articles by including one or more labels in the page's markup. This way, articles are grouped in categories and the category hierarchy forms a graph (multiple parents and cycles are allowed). Since users have the freedom to give any category name, the categorization can be very wide. Plamada and Volk [2012] already demonstrated the difficulty to use Wikipedia categories for the extraction of domain-specific articles from Wikipedia. So, categorization can help to assign a domain to an article, but it is not structured enough to do it in a straightforward manner.

For example, taking *Computer science* as a pseudo-root category to explore, and follow a path through one of its 13 subcategories we can observe that at depth 9 the domain has totally changed: *Computer science* → *Areas of computer science* → *Artificial intelligence* → *Human-computer interaction* → *Virtual reality* → *Virtual avatars* → *Fictional avatars* → *Fictional pharaohs* → *Pharaohs*.

Another feature of the graph is that a priori distant categories merge soon enough. Departing from *Literature* as pseudo-root category, one founds an intersection with *Computer science* in *Virtual reality* at depth 7. From there and below, subcategories would be equivalent for Computer Science and Literature. So, it is clear that if one aims at obtaining a collection of documents in the Computer Science domain, exploring the full graph is not an option.

We design a strategy to deal with the Wikipedia features shown above. First of all, we must identify a category with the domain we are interested in. Departing from this root category, the model performs a breadth-first search. Our stopping criterion is inspired by the Classification Tree-Breadth First Search [Cui et al., 2009]. The core idea is to score the explored categories in order to assess how likely it is that they actually belong to the desired area. In our approach, we make the naive assumption that a category belongs to the area if its title contains at least a word of the vocabulary of the domain. Nevertheless, many categories may exist that do not include any of the words in the vocabulary, so, at the end, we do not score every category individually but the whole level of categories at equal depth from the root.

The vocabulary of the domain is automatically built from the root category and it is composed by the tokens in its articles, but we only consider the most frequent 10% of tokens (after a standard pre-processing: tokenization, stop-words, filtering, and stemming). This value and the parameters used in this section were empirically chosen.

When exploring the graph we score each level by measuring the percentage of categories in it that are associated to the domain by means of this vocabulary. Experiments showed that those levels with at least a 50% of positive categories can be used to define the set of categories to be considered as representatives of the domain, so this is the stopping point in the search.

The corresponding Wikipedia dumps have been downloaded from the Wikimedia Downloads page[37] during January and February 2015 and preprocessed using the Java Wikipedia Library [Zesch et al., 2008]. For our study, we select two different root categories. With the final goal of improving translation engines for the Computer Science domain, we use *Computer science* and also *Science* as initial nodes. We expect *Science* to include *Computer science* and that would allow to gather larger corpora with a wider vocabulary, even less focused on the specific domain. Table 12 shows the number of articles and the maximum depth considered as part of the domain.

The collections of articles in the two languages constitute comparable corpora in the Computer Science and Science domains. From these collections, it is straightforward to select the parallel articles since they are connected via interlanguage links.[38]

For each pair of parallel articles, we estimate the similarity between all their pairs of cross-language sentences with different text similarity measures. We repeat the process for all the pairs of articles and rank the resulting sentence pairs according to its similarity. After defining a threshold for each measure, those sentence pairs with a similarity

[37] https://dumps.wikimedia.org

[38] An interlanguage link is a link from a page in one Wikipedia language to an *equivalent* page in another language.

category	language	#articles	depth	#parallel sentences
CS	English	155,533	7	577,428
	Spanish	29,634	6	
Science	English	785,642	8	3,847,381
	Spanish	820,949	6	

Table 12: Selected depth per category (CS = computer science), number of articles at the corresponding depth and number extracted parallel sentences.

higher than the threshold are extracted as parallel sentences. This is a non-supervised method that generates a noisy parallel corpus. The quality of the similarity measures will then affect the purity of the parallel corpus and, therefore, the quality of the translator. However, we do not need to be very restrictive with the measures here and still favor a large corpus, since the word alignment process in the SMT system can take care of part of the noise.

We compute similarities between pairs of sentences by means of cosine and length factor measures. The cosine similarity is calculated on three well-known characterizations in cross-language information retrieval and parallel corpora alignment: (i) character ngrams [McNamee and Mayfield, 2004] (ii) pseudo-cognates [Simard et al., 1992]; and (iii) word 1-grams, after translation into a common language, both from English to Spanish and vice versa. We add the (iv) length factor [Pouliquen et al., 2003] as an independent measure and as penalty (multiplicative factor) on the cosine similarity. For our experiments we use the corpora resulting after the union of the subcorpora extracted with each of the above-mentioned methods.

Evaluation in MT. We have built three MT systems for both the en→es and es→en translation directions. First, an SMT Baseline trained using the popular Europarl corpus for the translation model, and two systems trained on extra parallel corpora extracted from Wikipedia, in two ways: (i) only using articles about computer sciences (ii) union with sciences. For the monolingual LM, we interpolated the target side of the parallel corpora with the the news monolingual corpora released in the WMT 2012 Shared task on machine translation.

The development of all the systems was carried out using publicly available state-of-the-art tools: the mGIZA toolkit [Gao and Vogel, 2008], the SRILM toolkit [Stolcke, 2002] and the Moses decoder [Koehn et al., 2007]. More concretely, we followed the phrase-based approach with standard parameters: a maximum length of 80 tokens per sentence, translation probabilities in both directions with Good Turing discounting, word-based translation probabilities (lexical model, in both directions), a phrase length penalty and the target language model. The weights were adjusted using MERT tuning with n-best list of size 100. Development was done in Batch1.

Table 13 summarizes the results of the evaluation of the en→es and es→en on systems on QTLeap Batch2 test set (both questions and answers). First experiments on the Spanish–English language pair improve a baseline trained with the Europarl corpus in 3 points of BLEU also in the Computer Science domain. The models are preliminary, but we have been able to extract half million in-domain parallel sentences from the comparable corpora.

In the near future we want to improve the methodology to extract the sentences and

System	es→en	en→es
Europarl (baseline)	23.77	21.80
Europarl+WP(computer science)	25.46	**24.46**
Europarl+WP(computer science, science)	**25.91**	23.89

Table 13: The BLEU scores for Batch2 when using domain corpora from Wikipedia.

providing a tool to automatically extract comparable and parallel corpora from Wikipedia given a domain or, equivalently, a root category.

4.6 Summary of Experiment 5.4.3

The results gathered in this section show that specialized lexicons are indeed a key factor in a domain-specific setting like QTLeap. We have shown that dedicated heuristics to detect "fixed" NEs which should not be translated using a special "HideIT" rule-based module are helpful. We successfully extracted gazetteers from corpora and other resources, including the Wikipedia, closely related to Linked Open Data. Those gazetteers were used to improve both en→X and X→en translation directions. The improvements were higher for en→X (see Section 8). TM interpolation helped noticeably in both direction.

In addition, we have explored the translation of named-entities in news corpora using dedicated resources including Wikipedia and associated Linked Open Data resources, with very limited success. On the contrary, the use of Wikipedia to gather parallel domain-specific corpora provides notable improvements, which we would like to integrate in the QTLeap MT platforms for Pilot 3.

5 Experiments on coreference

Although coreference is not part of lexical semantics, it is one of the components that distinguishes our deep approach from the standard shallow approaches to MT. Therefore, we considered this report the best place to accommodate the experiments on using coreference in MT.

Even though addressing discourse-related issues is essential in MT, it tends to be commonly ignored and solved only implicitly. Coreference as one of the most important discourse phenomena plays a vital role especially when translating between two distant languages with different grammatical rules. For instance, while gender of English nouns is notional, keeping the male and female gender solely for persons, gender in Czech is morphological, with feminine, masculine and neuter evenly distributed also among the non-living objects. For both the language, a personal pronoun must agree in gender and number with the word it refers to (antecedent). However, since the systems of genders differ, one cannot just copy the gender of the English personal pronoun to Czech. This principle is nicely illustrated in Figure 5. Note that coreference-aware approach can be hardly replaced by a different approach for imposing target-language grammar rules. We addressed this problem for en→cs and en→nl translation in Section 5.1.

EN:	I bought a new $chair_{neut}$. I broke it_{neut}.
CS error:	Koupil jsem novou $židli_{fem}$. Zlomil jsem ho_{neut}.
CS ok:	Koupil jsem novou $židli_{fem}$. Zlomil jsem ji_{fem}.

Figure 5: An example of en→cs translation where it is necessary to resolve coreference to comply with the grammar rules of the target language. Treating the coreferential expressions *chair* and *it* independently produces a translation, where the genders of their Czech counterparts do not agree (CS error), unlike in the correct translation (CS ok).

Coreference appears between two mentions that refer to the same discourse entity. Thus, these mentions should be semantically compatible. In Section 5.2, we propose an additional method which combines the information gained by coreference resolution with WSD to ensure the semantic compatibility of coreference mentions. The experiments are conducted on en→bg.

5.1 Using coreference to impose target-language grammar rules

In these experiments we aimed at exploring how coreference information in the source language affects machine translation with TectoMT. We were especially interested in how to use coreference to impose grammar rules in the target language. Testing on en→cs and en→nl translation, we obtained coreference information by a coreference resolver (CR) for English and applied target-language rules which exploit this information.

To gain the coreference information, we integrated three coreference resolvers for English into TectoMT system, namely *BART* [Versley et al., 2008], the *Stanford CR* [Lee et al., 2013], and the *Treex CR* [Popel and Žabokrtský, 2010]. BART is a well-established modular toolkit for end-to-end coreference resolution. Its model is trained using the WEKA machine-learning toolkit [Witten and Frank, 2005]. The Stanford resolver is a state-of-the-art rule-based system organized in a sequence of sieves, which are the rules ordered by their decreasing precision. Treex CR is a rule-based coreference resolver and consists of several modules targeting personal, possessive and reflexive pronouns. It also

method	MT: BLEU		Intrinsic: F-score			
	Czech	Dutch	pers	poss	relat	total
Baseline	30.55	24.22	–	–	–	–
Treex-relat	31.08	**24.25**	–	–	73.64	–
Treex-relat+other	31.08	24.06	54.05	64.09	73.64	62.78
Stanford + Treex-relat	**31.10**	24.22	54.08	57.20	73.64	60.65
BART + Treex-relat	31.09	24.17	56.61	60.02	73.64	62.45

Table 14: BLEU scores (Batch2) of the TectoMT system for en→cs and en→nl translation using various CR systems, contrasted with their intrinsic evaluation measured by F-score.

targets relative pronouns, unlike the other two systems. Therefore, BART and Stanford CR were run in combination with the part of Treex CR aimed at relative pronouns (*Treex-relat*).

For both Czech and Dutch, we developed target-language rules exploiting the coreference information. The rules aim at imposing agreement in gender and number for personal, possessive, relative pronouns and their closest antecedents. Except for this, a specific pronoun can be enforced if the antecedent is of a predefined nature, e.g. Czech uses a word *svůj* for possessive pronouns that refer to a sentence subject.

The coreference-aware systems were evaluated on the Batch2a dataset. We compared it with the *Baseline* systems using no coreference-related rules and the systems provided only with the output of the Treex-relat resolver. Table 14 shows BLEU scores of all five configurations with respect to the target language. In addition, it presents an intrinsic evaluation of the CR – anaphora resolution F-scores measured on English parts of sections 20–21 of the Prague Czech-English Dependency Treebank [Hajič et al., 2012].

While there is a substantial BLEU difference between the Czech best system and the baseline, a very small improvement over the baseline of the best Dutch system has been observed. In addition, the improvements for Czech seem to be more consistent than for Dutch exhibiting a substantial drop for the Treex CR system. It may result from the fact that the en→cs system is more developed than the en→nl system, including the rules using coreference. The results of intrinsic evaluation show that out of all pronoun types, CR of relative pronouns is the most reliable. This is confirmed by gains of the *Treex-relat* system over the baseline, especially for Czech.

The bottleneck of using BART and Stanford CR is that they carry out their own linguistic analysis, including part-of-speech tagging and parsing. Higher time complexity together with marginal improvement compared to the Treex CR system convinced us to use the Treex CR system in Pilot 2 for en→cs translation. On the other hand, as we considered the improvements of the coreference-aware en→nl system marginal and not consistent enough, the Pilot 2 for this translation direction does not exploit coreference information.

Before we conclude, one apparent contradiction needs to be clarified. Whereas these experiments report a positive effect of CR for translation to Czech, the coreference-related parts in Section 9 of this deliverable and the D5.4 deliverable observe that the coreference tools run on the QTLeap corpus, formed by relatively short questions and answers, return too few coreference relations to affect the translation.

Surprisingly, both claims are true. The biggest improvement in Table 14 is observed for the Treex-relat system, which targets only relative pronouns. Relative pronouns are

not rare even in such short texts as the answers[39] from the QTLeap corpus. However, due to their very local scope many theories as well as CR tools do not consider them an anaphoric expression, e.g., the Stanford CR and BART. Stanford CR has been selected as the main coreference resolver for English in D5.4. A similarly working tool has been selected for Spanish, and possibly other languages. As a consequence, the observations of running these tools on the QTLeap data made in D5.4 and in Section 9 often do not take relative pronouns into account. For the future, we plan to study the situation in each language, and improve/adapt the coreference resolution tools as necessary.

The positive effect of CR in en→cs translation is in fact a synergy of several factors. Resolution of relative pronouns does not incur so many errors as for the other types of coreference. Furthermore, Czech relative pronouns need to agree in gender and number with its antecedent, which is hard to ensure by standard means, e.g., translation models. Moreover, as the relative pronoun is often a subject of the clause, the agreement is transferred also to the verb. Thus, it is no exception that correctly guessed pronoun causes a larger than a unigram match in BLEU score. Last but not least, TectoMT for en→cs has been developed for years, so the chance that a coreference-aware rule interferes with another rules is minimized.

On the other hand, in the cs→en translation of relative pronouns, the only decision to be made in English is whether a pronoun refers to a person. In that case, one has to use *who* instead of *which*. However, texts in the QTLeap domain almost never mention persons – *who* appeared only twice as a relative pronoun in English reference translations. Moreover, using *that*, regardless to what it refers to, is always correct. The situation is similar for personal pronouns. The distribution of genders in English and Czech over pronouns is different, Czech masculine and feminine pronoun can be used also to refer to many non-living entities. Such pronouns should be translated into the English pronouns *it*, or *they* in case the Czech pronoun is in plural. Again, due to lack of person entities in the QTLeap domain, *he* or *she* appear rarely, so translating a Czech personal pronoun always to *it*, or *they* works reasonably well. The same holds for possessive pronouns.

More information on this work can be found in Novák et al. [2015].

5.2 Using coreference for transferring semantic information

In these experiments, we consider a coreference chain as a mechanism for distribution of conceptual information within the text. First, coreference links between the mentions are obtained, e.g. between the mentions *London* and *the capital city*. Subsequently, the mentions' heads are annotated with the UKB system, which assigns them zero or more synset IDs, depending on the presence of the head term in WordNet. Moreover, every assigned synset IDs is accompanied with the score that UKB estimates using the context. For instance, the word *city* could belong to three possible concepts: `08524735-n` ("a large and densely populated urban area; may include several independent administrative districts"), `08540903-n` ("an incorporated administrative district established by state charter"), and `08226335-n` ("people living in a large densely populated municipality").

We envisage two ways of usage of this conceptual transfer. The first one is to gain new restrictions on the possible interpretation of the different terms. For instance, in the text "***XML*** *is used to aid the exchange of data.* ***It*** *makes it possible to define data in a clear way.*" the conceptual information can be transferred from the word *XML* to the pronoun *it*.

[39] Only answers, because the considered translation direction is from English.

However, in the current experiments we are using the coreference chains in the other way – to repair the semantic annotation of the text. Having two coreferential mentions in the text, we want them to agree on their semantic annotations. For example, in the text "*In the review tab, click where it says '**Language**', and then in 'Set **Proofing Language**'... select **Portuguese**.*" we want the three coreferring mentions in bold to have the same sense, that is, that each mention is annotated with a concept that is equal or more general to the conceptual annotation of the other. Therefore, if *Portuguese* is annotated as "Natural language that is spoken in Portugal (and some other countries)", we expect the two other terms to be annotated with the concept "Natural language".

We have performed these experiments for en→bg translation in the Moses system as an extension over the system described in section 3.2. Thus, we reused the model represented above with the addition of lexical semantics. After the semantic annotation with word senses and coreference by the UKB and the Corefgraph system,[40] respectively, we checked whether the heads of the coreferring chunks of text agree on conceptual information. If yes, we proceeded with the next coreference chain. If not, we tried to repair the word sense annotation by selecting an alternative annotation for one of the words. If it failed, the chain was skipped, leaving the word sense annotation as it was done initially. After this additional step, we applied the substitution of English words with a Bulgarian representative from the corresponding synset, as described in Section 3.2. The results are presented in Table 15, which shows the performance of the system using coreference with respect to the system which does not (cf. Table 6).

Source factors	Coreference	NIST	BLEU
ExpB (repres-lemma|form, lemma, PoS)	no	4.95	17.23
ExpB (repres-lemma|form, lemma, PoS)	yes	4.98	17.39

Table 15: NIST and BLEU scores for en→bg (Batch3a) using coreference over model reported in Table 6.

The results show a small improvement, and indication that coreference chains are useful to improve word sense disambiguation. We are hoping for larger gains in text which shows more coreference chains. For instance, in the parallel corpora used to train the MT system the document boundaries are not preserved, and thus coreference is limited to tokens in the same sentence. Thus, for the next pilot we envisage to add parallel texts which do contain document boundaries for the training, tuning and testing data. In this way we will exploit the full potential of the approach. In addition, we were not able to transfer directly the coreference information to the target language during decoding. We plan to work on that in the next pilot.

[40] http://ixa2.si.ehu.es/ixa-pipes/third-party-tools.html

6 Experiments on multiwords

In this Section, we present work conducted to analyse multiword expressions (MWEs) in the TectoMT system. We introduce a novel approach with several design goals. Firstly, it is automatic and wide-coverage, allowing construction of linguistic resources with a minimum of human effort, and requiring few or no external lexical resources or language-specific tools. Secondly, we place particular emphasis on a language-independent methodology, with obvious benefits for the QTLeap project and its many different languages. Finally, our method is domain-independent, meaning that the results we present here should be applicable also to translation applications not covered in the QTLeap project.

For the QTLeap project, we automatically build lists of multiword expressions for all QTLeap languages from raw text; this process is described in Sections 6.1 and 6.2. Section 6.3 describes an experiment which uses the English list inside the TectoMT system.

6.1 Acquisition of multiword expression candidate lists

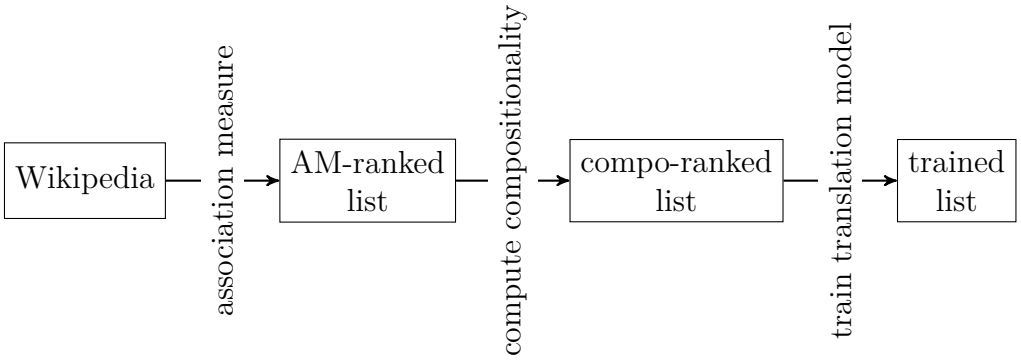


Figure 6: Simplified schematic of our multiword expression acquisition system.

We acquire lists of multiword expressions using a fully automatic unsupervised approach operating on large quantities of raw text. As shown in Figure 6, our acquisition method follows a pipeline model with re-ranking (Section 6.2) and filtering (Section 6.3) stages. We begin by constructing lists of MWE candidates using traditional association measures (AMs), which estimate how likely or unlikely is the combination of words in a particular multiword expression; association measures can be readily used to identify collocations (e.g., to detect that *salt and pepper* is much more common than *pepper and salt*). This list of candidates is then further processed to identify semantically idiosyncratic (non-compositional) MWEs, based on the intuition that non-compositional expressions should be more helpful for machine translation, since the meanings of such expressions are not predictable from their constituent words.

Association measures work by counting words and expressions, and benefit from a large amount of text. For our source material, we take the Wikipedia dumps from April 2015 in the various QTLeap languages. Wikipedia is a large textual resource available in multiple languages. It tends to be written in a uniform style, but covers a wide range of topics, and so may be regarded as a domain-independent corpus.

To keep our methodology wide-coverage and language-independent, we perform unrestricted identification of multiword expressions by collecting lexical co-occurrence statistics on all words in Wikipedia. This is a novel aspect of our current approach, since most

MWE acquisition research to date has focused on narrow linguistic categories, such as compound nouns, phrasal verbs, or light verb constructions. For the same reasons, we perform very little pre-processing of the text prior to MWE identification. We use the WikiExtractor tool[41] to retrieve only plain text from the Wikipedia dumps (discarding tables, images, formatting, and page links). We segment text into sentences, perform tokenization, and strip out URLs using simple regular expressions, and we remove all punctuation. Otherwise, we leave the text as it is (i.e., no POS-tagging, lemmatisation, case normalisation, or filtering out of numbers or symbols). Performing acquisition on unlemmatised text in this way may be useful for capturing the morphological and syntactic fixedness of some idiomatic MWEs.

We collect word frequency information on the Wikipedia text using the SRILM language modelling toolkit,[42] counting n-grams with n up to 3 (i.e., we treat MWEs as bigrams and trigrams). We use a threshold to prune the n-gram counts, discarding counts for unigrams, bigrams, and trigrams which are below a certain number. As indicated in Table 16, the threshold values vary by language, depending on how large the initial Wikipedia corpus is.

Language	Wikipedia Size (Tokens)	Vocabulary Threshold	Vocabulary Size (Types)
EN	2,789,274,024	20	1,078,567
DE	925,241,523	20	886,216
ES	565,343,796	15	470,682
NL	279,894,987	10	518,295
PT	262,901,712	10	377,646
CS	112,307,814	10	384,873
BG	54,318,288	5	318,435
EU	37,995,758	5	208,124

Table 16: Size of Wikipedia in the various QTLeap languages, showing varying word count thresholds, and resulting vocabulary sizes.

We then rank all bigrams and trigrams for each language in order of how strongly associated their constituent words are, using the Poisson collocation measure [Quasthoff and Wolff, 2002], which is identical up to a constant factor with the "log likelihood measure" introduced by Dunning [1993].[43] This is one of several association measures (AM) often used in the MWE literature; we chose it from a number of other such measures after a cross-lingual empirical evaluation, wherein we graded the association measures by how highly they ranked a set of known MWEs.

6.2 Compositionality ranking

We now take the top 10% from each association-measure-ranked list of MWE candidates, and re-rank the candidates in order of increasing compositionality. For this, we employ a

[41] https://github.com/bwbaugh/wikipedia-extractor

[42] http://www.speech.sri.com/projects/srilm/

[43] For this work, we introduce a variant of the Poisson measure which is suited to trigrams as well as bigrams. For a given MWE e consisting of words $w_1, w_2, \ldots, w_n$, with observed count $f(e)$, this is given by $\frac{f'(e) - f(e)\log f'(e) + \log[f(e)!]}{\log N}$, where N is the total number of unigrams in the corpus, and $f'(e)$ is the expected count of a MWE: $f'(e) = N^{1-n} \prod_i^n f(w_i)$.

method based on the work of Salehi et al. [2015], who recently obtained state-of-the-art results on multiword compositionality ranking. The method makes use of word embeddings constructed using the word2vec[44] software [Mikolov et al., 2013a]. We build a vector representation for every word in the vocabulary, as well as for every MWE candidate, using the extracted Wikipedia text by greedy string search-and-replace of all occurrences of MWE candidates, replacing each of these with a single words-with-spaces token.

A drawback of our acquisition method is that the greedy string rewriting cannot deterministically handle n-grams that overlap with other n-grams. Thus, in order to perform this string rewriting, we sort the MWE candidates in the AM-ranked list into 10 batches, such that no two MWE candidates in each batch overlap with each other (i.e., for all e_1, e_2 in each batch, e_1 is neither a substring nor a superstring of e_2, and there is no prefix (or suffix) of e_1 which is a suffix (or prefix) of e_2). This sorting is performed greedily by processing MWE candidates in order of decreasing association measure, and assigning each MWE candidate to the first batch found which preserves this property. MWE candidates which cannot be assigned to one of these 10 batches are discarded. Therefore, n-grams containing very frequent words can be discarded in this step because they overlap with a large number of other n-grams. On our lists, we observe the discarding of between 2–10% of MWE candidates.

Each batch then produces a word embedding model for all words in the vocabulary, and some subset of MWE candidates. To compute compositionality, we compute the cosine similarities of the vector representation of each MWE candidate with the vector representations of its constituent words, and take the arithmetic mean. In performing this calculation, we do not compute the similarity of MWE candidates with any stop words that they may contain.[45] The MWE candidates from all batches, with their associated compositionality scores, are then recombined and sorted to produce a single compositionality-sorted list.

This procedure is unable to compute compositionality scores for MWE candidates that contain words not found in any other contexts (e.g., the name "Neil deGrasse Tyson", as the word "deGrasse" is not found in the English Wikipedia, except inside this MWE candidate). These candidates are assigned an arbitrary, large negative compositionality score (-2, lower than the lowest item in the compositionality-ranked list).

6.3 Initial experiments incorporating multiwords in TectoMT

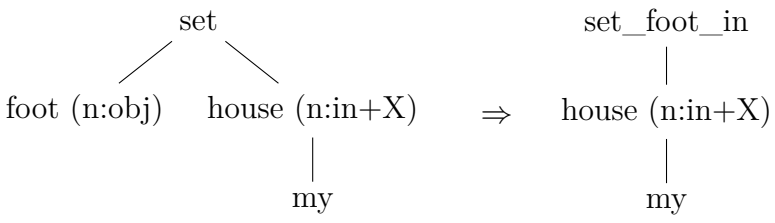


Figure 7: Tectogrammatical reduction of multiple t-nodes (representing the non-compositional multiword expression "set foot in") into a single composite t-node.

[44] https://code.google.com/p/word2vec/

[45] For this work, we define the stop words in each language to be the 50 most frequent words in the vocabulary.

In this Section, we describe an experiment incorporating the list of non-compositional English MWEs into the TectoMT system.

The basic aim of this work is to collapse multiword expressions in the tectogrammatical dependency graph into a single composite *t*-node; we expect that a single *t*-node should be easier to translate than multiple connected nodes that express an idiosyncratic meaning. As illustrated in Figure 7, this procedure consists of altering the lemma of the topmost *t*-node to represent the MWE as a word-with-spaces, deleting dependent multiword nodes, and rearranging arguments to the MWE so that they depend on the new composite node.

In principle, this analysis can be performed in both the source and target languages, with the translation model learning the mapping between them. In practice, the collapsing of the MWEs into composite *t*-nodes must be reversible to support generation in the target language. As there are several design choices to be made with respect to generation, we opted to first establish the validity of our approach with a simpler experiment, in which we analyse multiwords only in the source language. Note that this paradigm will only work for multiwords which can be translated into single lexical nodes in the target language; MWEs which are translated by other multiwords will result in translation failures (i.e., insertion or deletion errors). However, we expect that such errors will happen relatively infrequently (cf. Uresova et al. [2013], who found in the Parallel Czech-English Dependency Treebank that most verbal MWEs are not translated by other MWEs).

Modifying the TectoMT system to analyse MWEs in the source language requires building new translation models, a time-intensive process, so we conduct our experiments with the en→es pipeline, which we found to be less computationally complex than the other QTLeap languages. Thus, these experiments only make use of the English list of non-compositional MWEs.

In preparation for this experiment, we perform some final filtering to the English list, removing several of the more common errors that we observed using a simple pattern-based filter (e.g., discarding those candidates which begin or end with a conjunction or some form of the copula). We also discard some MWE candidates which are superstrings or substrings of another, less compositional MWE candidate, when the two candidates have similar word embedding vectors. This results in the removal of around 9% of the candidates from the list, leaving us with 551,253 English MWE candidates with compositionality value less than 0.5.

Our MWE candidate lists are constructed using *n*-grams, and contain many candidates that are not syntactic units, such as "to found the" and "no husband present". Therefore, our analysis operates after parsing has taken place, and we only match as MWEs sets of nodes in the dependency trees that are "treelets" (we match only strings on nodes which are fully connected to each other by dependency relations).

We identify multiwords in the parsed text greedily by matching on word forms in the *a*-trees (analytical layer trees, see D2.8). In this search, multiwords with lower compositionality scores are preferred, and ties are resolved arbitrarily by taking the leftmost match. Figure 7 shows the reduction performed in the analysis of a successfully matched MWE instance.[46] We identify the *t*-nodes corresponding to the matched *a*-nodes, and "reduce" the MWE treelet.

We record all MWEs seen during training, and use only this list for analysis during testing, to ensure that no MWEs are reduced for which the trained translation model has

[46] In this example, the preposition "in" has been encoded in the formeme of the *t*-node under it ("house") by the TectoMT system, but our analysis will still find this treelet because it can find "set" and "foot".

not learned any translations (which would create new out-of-vocabulary items). This has the effect of filtering our MWE candidate lists, so that, at test time, only those expressions found in the translation training corpus are used to analyse the test data.

To investigate the effect of this filtering, we train translation models on two corpora. The Europarl corpus (50 million words, 2 million sentences) is the standard parallel data used for training en→es models. We also constructed an "in-domain" corpus, consisting of about 25 million words in 1.25 million sentences, from the sentence-aligned EN-ES parallel text in the "KDE" and "OpenOffice" files, and half of the "commoncrawl" file. This material is predominantly technical in nature, and represents a better domain overlap with the QTLeap test set. We manipulate the compositionality value as an independent variable, using a threshold θ to control the number and compositionality of MWEs that are analysed in the source (English) text. For example, with $\theta = 0.1$ we restrict the MWE candidate list to contain only those items whose compositionality score is less than 0.1. BLEU score results of these experiments are shown in Table 17,[47] and Table 18 shows the counts of MWEs found during training and testing.

	Europarl	In Domain
No MWEs	20.24	26.00
MWEs, $\theta = 0.1$	–	26.46 ***
MWEs, $\theta = 0.2$	20.19	26.43 **
MWEs, $\theta = 0.3$	–	26.08
MWEs, $\theta = 0.4$	–	25.48
MWEs, $\theta = 0.5$	19.39	24.55

Table 17: Summary of BLEU scores for en→es translation models trained on Europarl and in-domain text, tested on Batch2a. Statistical significance with respect to the baseline: ** $p < 0.01$, *** $p < 0.001$.

The experiments demonstrate that source-only analysis of automatically acquired MWEs improves translation quality for this language pair (+0.46 BLEU points). The improvement is only seen for the models built with the in-domain text, which we take as an indication that our approach is sensitive to the domain of the training data, for the aforementioned reason that the training corpus acts as a filter on the MWE lists. Europarl, as the larger corpus, is a less strict filter than the in-domain text (cf. the larger MWE counts for Europarl under the "Training" columns in Table 18); however, the MWEs trained using the Europarl corpus seem to be a poorer thematic fit for the QTLeap test data than those found in the in-domain text (cf. the smaller MWE counts under the "Test" columns).

This evaluation paradigm is sensitive to the compositionality of the MWEs, as the greatest improvements over the baseline are seen with small values of θ; including more compositional MWEs ($\theta > 0.3$) eventually reduces BLEU scores below the baseline. This effect is expected, because it is likely that composite t-nodes representing compositional MWEs cannot be adequately translated by single lexemes. Finally, we note that the in-domain model with $\theta = 0.1$ outperforms the baseline, even though the model does not analyse any MWEs in the test set. This suggests that our approach may not only

[47] These experiments were conducted using an early development version of the Pilot 2 software; for this reason, the Europarl baseline number shown here differs from the Pilot2-minus-LS value shown in Table 23.

Experiment	Training		Test	
	Types	Tokens	Types	Tokens
Europarl				
$\theta = 0.2$	5,020	174,015	7	8
$\theta = 0.5$	90,133	2,808,015	220	331
In Domain				
$\theta = 0.1$	837	4,593	0	0
$\theta = 0.2$	3,576	19,586	11	14
$\theta = 0.3$	12,333	67,709	52	95
$\theta = 0.4$	32,126	160,828	138	234
$\theta = 0.5$	61,657	303,724	293	480

Table 18: Number of MWE items found during training and testing: en→es translation models trained on Europarl and in-domain text, tested on Batch2a.

improve results on text containing MWEs, but may also improve the overall quality of the translation model.

This experiment validates our treatment of MWEs, demonstrating a positive effect from adding MWEs to the TectoMT system, despite our very simple analysis architecture. In future, we will analyse MWEs in both the source and target languages. This will have the useful side-effect of permitting a degree of non-isomorphic transfer not possible using the current architecture: A system supporting generation of MWEs in the target language will be able to additionally translate single lexemes to multiwords, and multiwords to multiwords, resulting in fewer translation errors where the current MWE-to-single-lexeme model fails. Furthermore, such a system should exhibit less sensitivity to the value of the compositionality threshold.

7 Experiments with Qualitative MT on en→de

The experiments for German undertaken in the development of Pilot 2 have been restricted to the translation direction en→de. Although German is not part of the working package of this deliverable, we take the opportunity to report a series of experiments on different components of the hybrid architecture which are related to the experiments in the previous section.

First, we have experimented with special lexicons on the rule-based component of the QTLeap Qualitative MT platform. To this end, we have converted the lexicon based on Wikipedia entries produced within the project (see Section 4.2) so that it could be imported as a special lexicon. Likewise, we have converted the Microsoft Terminology. The number of collected gazetteer entries are shown in Table 8. We also manually coded about 230 frequent terms from the test batch (Batch2a) to assess the effect of an "oracle". For all experiments, the results in terms of BLEU were marginal. Testing on Batch2a, the baseline Lucy system had a BLEU score of 26.08. Inclusion of the Gazetteers led to BLEU scores of 26.16 (Wikipedia Gazetteer) and 26.38 (Microsoft Gazetteer), respectively. Manual coding of the most frequent unknowns also led to a BLEU of 26.38. Table 19 summarizes the results.

Experiment	BLEU
Baseline	26.08
Wikipedia Gazetteer	26.16
Microsoft Gazetteer	26.38
Manual coding	26.38

Table 19: BLEU scores of German rule-based component using different Gazetteers.

Manual inspection of the outputs of the Lucy baseline system and the system with the extension by the Microsoft terminology revealed improvements due to the better handling of terminology. However, some errors have been introduced as well, e.g., the new occurrences of untranslated terms as the Gazetteers include English-German entries such as *Access – Access* for the respective Microsoft product that have a negative effect on the translation of the word *Access* in sentences like "*Access your hard drive…*". Figure 8 illustrates improving and worsening segments.

Second, we experimented in our serial system combination where the transfer-based output is automatically post-edited by an SMT component with the introduction of pseudo-senses in cases where the transfer-based component could not disambiguate. In Pilot 1, we always chose the first sense (alternative) in such cases, but for Pilot 2, we created pseudo senses like "table+whiteboard" for the translation of the German *Tafel* and trained the statistical system on the respective senses. The effects in terms of BLEU were marginal, however, so that we decided to not pursue this line of experiments further.

Third, we extended our statistical component with WSD based on the model for English described in [Weißenborn et al., 2015]. We ran offline experiments where the senses are used as alternative paths in Moses with four settings:

1. Baseline
2. Sense → Word
3. Word → word, Sense → word (alt path)

Source	You can download the MEO Drive from Google Play .
Reference	Sie können die MEO - Drive von Google Play herunterladen .
Pilot 2.00	Sie können das MEO - Laufwerk von Google - Spiel herunterladen .
Pilot 2.01	Sie können das MEO - Laufwerk von Google Play herunterladen .

Source	Go to the menu File > Open File . . .
Reference	Gehen Sie auf Datei > Datei öffnen . . .
Pilot 2.00	Gehen Sie zur Menü - Datei - > - Offen - Datei . . .
Pilot 2.01	Gehen Sie zum Menü File > Eröffnen File . . .

Figure 8: Examples of improving and worsening transfer-based output without (Pilot 2.0) and with (Pilot 2.01) Microsoft Gazetteer.

4. Sense → word, Word → word (alt path)

The senses used for training and decoding by Moses are estimated based on the disambiguation analysis on the sentence level by choosing the best ranked sense from the WSD system. The produced WSD labels are concatenated with the respective base word forms. In the alternative path, non-annotated input is used. The alternative path allows for decoding phrases when there are no WSD labels or the decoder cannot form a translation with a good probability. Due to the high complexity of the WSD annotation, this model was trained on less data than the respective phrase-based models for Pilot 0 and Pilot 1.

We experimented with all available English QTLeap corpora including questions. Table 20 provides selected results. Precision and Recall have been computed on the n-gram level using `analysis.perl` from the Moses toolkit. Results were promising as we got 1 BLEU score improvement for Batch3q and 0.4 BLEU score improvement for Batch1q on experimental setting 4 (other settings did not improve over the baseline, and are thus not shown in the Table). The table does not show results for answers, as no improvement was detected. In fact, the best results were achieved on questions rather than answers. The improvement on the news datasets was non-existent or small. For Pilot 2, we set up a REST server so that the disambiguation can happen at real time.

	Exp. No.	BLEU	METEOR	precision	recall	F1
batch1q	1	19.51	0.4502	0.604	0.629	0.617
	4	**19.97**	**0.4572**	**0.615**	**0.636**	**0.626**
batch2q	1	29.47	0.5257	0.684	0.698	0.691
	4	27.91	0.5186	**0.692**	0.698	**0.695**
batch3q	1	20.82	0.4268	0.590	0.605	0.598
	4	**21.82**	**0.4416**	**0.621**	**0.621**	**0.621**
news2012	1	18.83	0.4282	0.590	0.582	0.586
	4	17.85	0.4242	0.586	0.577	0.582
news2013	1	17.41	0.4043	0.569	0.555	0.562
	4	16.78	0.4024	**0.574**	0.554	**0.563**

Table 20: Experiment results on WSD for German.

8 Results of lexical semantics on Pilot 2 systems

In this Section, we describe the results of the techniques that were successful in the experiments mentioned above. Table 21 summarizes the experiments presented in this deliverable, specifying which ones were successful and which ones have been integrated in Pilot 2.

Exper	Sec.	MT	LS technique	Languages	Datasets	OK	P2
5.4.1	2.1	TectoMT	WSD (UKB)	en→pt	QTa		
	2.2	TectoMT	WSD (UKB)	en→nl	QTa		
5.4.2	3.1	TectoMT	WSD (UKB)	en→pt	QTa	Yes	Yes
	3.2	DFMT	WSD (UKB)	en↔bg	QTa,QTq		
	3.3	Moses	WSD (SStagger)	en→es	QTa,QTq,NW	Yes	
	3.4	TectoMT	Embeddings	en→cs	QTa,QTq,NW	Ong.	
5.4.3	4.1	TectoMT	Fixed entities	en↔{cs,es,eu,nl,pt}	QTa,QTq	Yes	Yes
	4.2	TectoMT	Gazetteers	en↔{cs,es,eu,nl,pt}	QTa,QTq	Yes	Yes
	4.3	TectoMT	TM interpolation	en↔{cs,es,eu,nl,pt}	QTa,QTq	Yes	Yes
	4.4	Moses	NERC	en→es	NW		
	4.5	Moses	Domain corpora	en↔es	QTa,QTq	Yes	
Coref	5.1	TectoMT	Coreference	en→{cs,nl}	QTa	Yes	Yes
	5.2	DFMT	Coreference	en→bg	QTa		
MWE	6.1	TectoMT	MWE source-only	en→es	QTa	Yes	
German	7	QMT	Gazetteer	en→de	QTa	Yes	Yes
	7	QMT	WSD (own)	en→de	QTa		

Table 21: Summary of experiments, including success and integration in Pilot 2. Columns stand for the following. **MT** for the MT platform, with the following values: TectoMT for the QTLeap TectoMT platform, DFTM for the QTLeap Deep Factored TM platform, QMT for the QTLeap Qualitative system combination platform. **Datasets**: QTa for QTLeap answers, QTq for QTLeap queries, NW for WMT news. **OK** for successful improvement over baseline, where Ong. is used for ongoing experiments. **P2** for use in Pilot 2

The successful components employing lexical semantics that were introduced in the previous sections are evaluated within the Pilot 2 systems, with the exception of two of the experiments (domain corpora and multiword expressions) which will be integrated in the next Pilot. The results in terms of BLEU scores are shown in Tables 22 and 23 for translation to English and from English, respectively. All the systems were evaluated on the Batch 2 dataset, which is used as a development dataset. Note that Batch 3 was reserved for the final testing Pilot 2, as described in Deliverable D2.8.

The tables report several baselines, including the scores of the Pilot 0[48] and the Pilot 1 systems. The row denoted as Pilot2-minus-LS shows BLEU scores of the Pilot 2 systems, if all the lexical semantics components are switched off. Each row that follows presents one of the lexical semantics component and what is the effect of switching on just this single component in the Pilot2-minus-LS system. The "Δ total LS" row shows the effect of switching on all the components (with two exceptions mentioned in the next paragraph). The "Δ total LS" is usually not a sum of the differences for individual components, as

[48]The Pilot 0 results for en→es and es→en reported here are Pilot 0-comparable, that is Pilot 0 trained on Europarl only, so it can be fairly compared with Pilot1 and Pilot 2, which are also trained on Europarl only.

the effects of these components may overlap. The final performance of the full Pilot 2 systems can be found in the last row of the tables.

Note that Pilot 2 did not activate all components (rows) reported in the table, as there are two exceptions: 1) Using a gazetteer in the nl→en translation deteriorate noticeably the score, so we decided to deactivate the gazetteer module in full Pilot 2 for nl→en. 2) The effect of +synset(node,sibling) in en→pt is positive, but it is mutually exclusive with +synset&supersense(node,parent), which brings a higher improvement, and therefore the later was kept.

system	cs→en	es→en	eu→en	nl→en	pt→en
Pilot0	26.44	39.30	25.29	36.45	22.59
Pilot1	26.81	16.05	4.75	34.46	10.14
Pilot2-minus-LS	27.78	26.21	13.30	44.01	11.94
Δ "fixed" entities (HideIT)	−0.01	+0.01	+0.03	+0.00	+0.01
Δ specialized lexicons (gazetteers)	+0.77	+0.62	+0.00	−0.09	+0.02
Δ adaptation by TM interpolation	+1.67	+0.42	+0.71	+1.91	+1.50
Δ total LS	+2.50	+0.94	+0.77	+1.92	+1.57
full Pilot2	30.28	27.15	14.07	45.93	13.51

Table 22: Translations to English (Batch2q). Effect of various lexical semantic modules on BLEU performance.

system	en→cs	en→es	en→eu	en→nl	en→pt
Pilot0	31.07	25.11	28.37	32.94	19.36
Pilot1	30.68	16.92	14.39	23.10	19.34
Pilot2-minus-LS	28.07	26.25	20.87	23.38	19.82
Δ +synset(node,sibling)					+0.20
Δ +synset&supersense(node,parent)					+0.25
Δ "fixed" entities (HideIT)	+0.84	+0.46	+0.56	+0.48	+0.34
Δ specialized lexicons (gazetteers)	+3.49	+3.19	+0.91	+1.49	+0.94
Δ adaptation by TM interpolation	+0.74	+5.10	+0.06	+0.75	+1.98
Δ total LS	+4.97	+7.91	+1.46	+2.45	+2.60
full Pilot2	33.04	34.16	22.33	25.83	22.42

Table 23: Translations from English (Batch2a). Effect of various lexical semantic modules on BLEU performance.

9 Evaluation of LRTs

D5.6 reported the Named-Entity Disambiguation (NED), Word Sense Disambiguation (WSD) and coreference tools for Basque, Czech and Portuguese. This section provides the evaluation of these tools in standard evaluation corpora, as well as the domain-specific QTLeap corpus. This section should be read in conjunction with Section 7 of deliverable 5.4, which reports the evaluation of the rest of the tools.

9.1 Basque

9.1.1 NED

The ixa-pipe-ned-ukb module for Basque has been evaluated on the publicly available EDIEC (Basque Disambiguated Named Entities Corpus) dataset.[49] This dataset is a corpus of 1032 text documents with manually disambiguated NEs [Fernandez et al., 2011]. The documents are pieces of news of the 2002 year edition of the Euskaldunon Egunkaria newspaper. We obtained a performance of 90.21 in precision, 87.90 in recall and 87.90 in F1 [Pérez de Viñaspre, 2015].

9.1.2 WSD

The ixa-pipe-wsd-ukb module for Basque has been evaluated on the publicly available EPEC-EuSemcor dataset.[50] This dataset is a Basque SemCor corpus, that is, a Basque sense-tagged corpus, which comprises a set of occurrences in the Basque EPEC corpus [Aduriz et al., 2006], which has been annotated with Basque WordNet v1.6 senses [Pociello et al., 2011]. More specifically, it contains 42,615 occurrences of nouns manually annotated, corresponding to the 407 most frequent Basque nouns. We obtained a performance of 56.5 in precision, 56.3 in recall and 56.4 in F1.

9.1.3 Coreference

The ixa-pipe-coref-eu module has been evaluated on the publicly available EPEC-KORREF dataset.[51] This dataset is a corpus of Basque text documents with manually annotated mentions and coreference chains, which consists of 46,383 words that correspond to 12,792 mentions. The document collection is a subpart of the Basque EPEC corpus (the Reference Corpus for the Processing of Basque) [Aduriz et al., 2006], which is a 300,000 word sample collection of news published in Euskaldunon Egunkaria, a Basque language newspaper. Our best system score 53.67 CONLL F1, 5 points above baseline (48.67) [Soraluze et al., 2015]. The baseline system is a copy of the original Stanford Deterministic Coreference Resolution System [Lee et al., 2013], taking as input only the output of the Basque linguistic processors and translated static lists, and our best system modifies and adds some sieves taking advantage of the morphosyntactic features of Basque.

9.1.4 Domain evaluation

NED For 869 of the total NERC mentions in the QTLeap corpus that we examined, the named entity linking module was able to find a link to DBPedia resources for 252 mentions.

[49] http://ixa2.si.ehu.es/ediec/ediec_v1.0.tgz
[50] http://ixa2.si.ehu.es/mcr/EuSemcor.v1.0/EuSemcor_v1.0.tgz
[51] http://ixa2.si.ehu.es/epec-koref/epec-koref_v1.0.tgz

Domain-specific entities were mostly correct, and it seems that the tool performed at the expected level. For instance, *Sareko* and *Facebook* were linked to `http://eu.dbpedia.org/resource/Internet` and `http://eu.dbpedia.org/resource/Facebook`, respectively. Even domain-specific products such as *Java* and *MB* were correctly linked to `http://eu.dbpedia.org/resource/Java_(programazio_lengoaia)` and `http://eu.dbpedia.org/resource/Megabyte`. We see, however, some room for improvement with cases such as *PS* which was incorrectly linked to the French Socialist Party `http://eu.dbpedia.org/resource/Frantziako_Alderdi_Sozialista`.

WSD Word disambiguation was performed for 24,691 tokens out of a total of 53,239 present in the Batch 1 and Batch 2 of the QTLeap corpus. This means that 46.38% of the tokens were linked to WordNet and were thus disambiguated. Many disambiguations were correct, and we do not see any performance loss from the expected values. Such is the case of the domain-specific noun *menu*, for instance, which was linked to the synset `30-06493392-n` with a confidence of 0.30, specifying "computer menu". A number of incorrect cases were found, such as domain-specific *mouse*, for instance, which was linked to the `30-02330245-n` with a confidence of 0.52, referring to the animal, instead of the correct synset `30-03793489-n`, with confidence 0.48, which is the specific synset for the IT domain.

Coreference As pointed out in Section 7.4.7.6 in D5.4, the QTLeap use scenario is quite peculiar from a coreference point of view. The user-machine interactions generally consist of one user question and one answer. The answer usually consists of one sentence, but occasionally a few short sentences are displayed. In this context, the number of coreferences present in the texts is low.

9.2 Czech

9.2.1 NED

As pointed out in Section 7.3.3 in D5.4, NameTag NERC tool was used for named entity recognition subtask. Its F1 measure on the test portion of Czech Named Entity Corpus 2.045 [Ševčíková et al., 2014] is 80.30% for the coarse-grained 7-classes classification and 77.22% for the fine-grained 42-classes classification. Nowadays there does not exist any publicly available test set for the evaluation of NED, but we evaluated the obtained results manually in Section 9.2.4.

9.2.2 WSD

As described in Section 5.4.2 in D5.6, two different approaches were used for WSD. The approach based on work [Dušek et al., 2015] was evaluated on Prague Czech-English Dependency Treebank 2.0 [Hajič et al., 2012] and showed 80.47% F1 score. For the second approach to WSD, there is no publicly available test set. Therefore, we provide manual evaluation of results obtained in Section 9.2.4.

9.2.3 Coreference

Coreference resolver for Czech has been evaluated separately for three different classes of anaphors: relative pronouns, a joint group of subject zeros, personal, and possessive

pronouns (all in 3rd person), and noun phrases. For each anaphor, F-scores of finding any of its antecedents were measured. The scores were calculated on the evaluation set of Prague Dependency Treebank 3.0 [Bejček et al., 2013]. The rule-based resolver for relative pronouns performs with the F-score 67.04%. The resolvers for the other two classes follow the machine learning approach. While the resolver for pronouns and zeros achieved the F-score 50.28%, resolution on noun phrases performs the worse, with the F-score 44.40%. The order of how the resolvers ranked correlates with the general linguistic feeling on complexity of finding an antecedent for individual anaphor classes.

9.2.4 Domain evaluation

NED Domain evaluation of NameTag results in named-entity recognition subtask was presented in Section 7.3.5.3 of D5.4. For 1,715 of total recognized entities, the named-entity linking was able to find a link to 572 DBpedia resources. Domain-specific entities were mostly correct, and it seems that the tool performed at the expected level. For instance, the terms *Gmail* and *Skype* (in any of its inflectional forms) were linked to `http://dbpedia.org/resource/Gmail` and `http://dbpedia.org/resource/Skype`, respectively. There is, however, some room for improvement in cases when NameTag marks some numbers as possible NEs and then the linking algorithm assigns a link to the corresponding page on Wikipedia. Those pages tend to refer to dates or numerical values, which usually does not make much sense in the IT domain.

WSD Word disambiguation was performed for 11,060 tokens out of a total of 71,061 present in the Batch 1 and Batch 2 of the QTLeap corpus. This means that 15.5% of the tokens were linked to Valency Lexicon [Urešová, 2011] and were thus disambiguated.

The second approach to WSD, described in Section 5.4.2 of D5.6, was applied to Europarl parallel corpus. The performance seems reasonable, for example, it produced mappings for words *zasedání* and *rozprava* to synsets `30-07145508-n` and `30-07140978-n`, respectively.

Coreference We ran the resolvers on the Batch2q dataset, consisting of 1000 questions. A coreference relation was found only in 65 sentences in case of relative pronouns, and in 49 sentences in case of zeros, personal, and possessive pronouns. This accords with what has been observed for English in D5.4 and for Basque and Portuguese (cf. Sections 9.1.4 and 9.3.4).

Moreover, the observations made in Section 5.1 show that Czech pronouns often carry more information than English, which is emphasized by the nature of the QTLeap domain, lacking person entities. Therefore, coreference would bring no additional information to be exploited in the cs→en translation.

9.3 Portuguese

9.3.1 NED

The NED-PT tool (originally described in section 5.5.1 of D5.6) has been evaluated using a gold-standard, NE-annotated version of the CINTIL International Corpus of Portuguese [Barreto et al., 2006]. The original corpus comprises approximately 1 million tokens manually annotated with lemmas, part-of-speech, inflection, and NEs, and contains data from both written sources and transcriptions of spoken Portuguese. The annotation

of NEs within the corpus with their corresponding Portuguese Wikipedia entries from DBpedia [Lehmann et al., 2012] was completed using version 1.3 of the brat web-based annotation tool [Stenetorp et al., 2012a,b].

Of the 26,371 NEs in the CINTIL corpus, 16,120 have been manually disambiguated. 12,160 of these 16,120 manually disambiguated entities are also automatically disambiguated by NED-PT, from a total of 16,486 tagged by the program. We thus define recall as the number of entities with the same DBpedia entry assigned by both NED-PT and the human annotator, divided by the number of entities manually disambiguated (16,120). NED-PT assigned the same DBpedia entry to the entity as was chosen by the annotator for 9484 of the 12,160 entities for which entities were assigned both manually and automatically, giving a precision of 77.99%, recall of 58.83% and F1 of 67.07%. In counting the accurate results (same DBpedia entry assigned to the entity by both the annotator and by NED-PT) we take into account those for which the assigned DBpedia entries may appear different at first glance, but in reality redirect either to or from each other in the DBpedia and Portuguese Wikipedia hierarchies.

9.3.2 WSD

Like NED-PT, the WSD-PT tool (originally described in section 5.5.2 of D5.6) has also been evaluated using a gold-standard, sense-annotated version of the CINTIL International Corpus of Portuguese [Barreto et al., 2006]. Of the 700,000 tokens we used from the written part of the corpus, 193,443 are open class words. The word-sense annotated version of the corpus was manually annotated with ILIs from the Portuguese MultiWordNet (approximately 19,700 verified synsets) [MultiWordNet, n.d.] using the LX-SenseAnnotator tool [Neale et al., 2015].

Of the 193,443 open class words in the CINTIL corpus, 45,502 have been manually disambiguated. 45,386 of these 45,502 manually disambiguated words are also automatically disambiguated by WSD-PT, from a total of 59,190 tagged by the algorithm (human annotators may have chosen not to disambiguate certain words for some reason, but the UKB algorithm will always assign something from the options available to it). We thus define recall as the number of words with the same sense assigned by UKB and the human annotator, divided by the number of words manually disambiguated (45,502). WSD-PT assigned the same sense to the word as was chosen by the annotator for 29,540 of the 45,386 words for which a sense was assigned both manually and automatically, giving a precision of 65.09%, recall of 64.92% and F1 of 65.00%.

9.3.3 Coreference

Our goal is to evaluate (and train) the Portuguese Coreference tool (originally described in section 5.5.3 of D5.6) using the Summ-it Corpus (v3.0) [Collovini et al., 2007], a corpus of coreference for Portuguese constructed from 50 news texts from the 'caderno de Ciência da Folha de São Paulo'. The corpus has been automatically annotated with morphosyntactic information and then manually with information about coreference between nominal phrases and about rhetorical relations. So far, dealing with inconsistencies in Summ-it has been highly problematic, both for training the tool (as described in section 5.5.3 of D5.6) and for later evaluation. We plan to make some fixes to the corpus to try and make it more usable for our needs, but for now use it as a guide for estimation.

For 316,000 sentences of the Portuguese side of Europarl (10 million tokens), the Portuguese Coreference tool was able to find 727,142 markable pairs, from which 22,984

(3.16%) are coreferent. While we would expect that far fewer of the possible pairs of markables in a given document are coreferent than not, this number still seems low. Attempting to run the tool over the Summ-it corpus (819 sentences), the tool was able to find 17,901 markable pairs, of which just 270 are marked as coreferent. Our work with the Summ-it corpus so far suggests to us that there are far more markable pairs to be found within the corpus, and that a far higher percentage of those pairs are coreferent.

One possible cause for the tool's low recall of markable pairs could be inconsistencies between the dependency-parsed and consitutency-parsed inputs over which the tool runs, leading in many cases to the failure of the head-finding heuristics on which the tool makes decisions. Further work on the Portuguese Coreference tool is required, therefore, to ensure that it captures more of the markable pairs present in given texts and can more accurately decide whether they are likely to be coreferent or not.

9.3.4 Domain evaluation

NED NED-PT was used to process Batch 1 and 2 of the QTLeap corpus (2000 questions and 2000 answers). From 3,799 entities found, 1,868 (49.17%) were disambiguated and linked to their Portuguese Wikipedia entries via DBpedia. Domain-specific entities are mostly correct, suggesting that the tool performs at the expected level. For example, *ISP* was linked to the Portuguese `http://pt.dbpedia.org/resource/Fornecedor_de_acesso_à_Internet` and subsequently `http://dbpedia.org/resource/Internet_service_provider`, and *écran* linked to the Portuguese `http://pt.dbpedia.org/resource/Monitor_de_vídeo` and subsequently `http://dbpedia.org/resource/Electronic_visual_display`. We do however notice some incorrect cases, most notably where the desired link for a particular entity does not share a Wikipedia/DBpedia entry in both Portuguese and English – for example *reiniciar* was linked to the Portuguese `http://pt.dbpedia.org/resource/Reboot_(ficção)` and subsequently `http://dbpedia.org/resource/Reboot_(fiction)`, denoting *reboot* in the sense of book and movie franchises. The desired `http://dbpedia.org/resource/Reboot_(computing)` does not have an equivalent entry in Portuguese, and so was not found.

WSD Processing Batch 1 and 2 of the QTLeap using WSD-PT, 6,115 (20.40%) terms were disambiguated from a total of 29,895 open-class words. The low recall seen here is likely to be a result of the lack of domain-specific terms in the Portuguese MultiWordNet, over which WSD-PT performs WSD. Many of the domain-specific terms that were evaluated appear to be correct, suggesting that the tool is performing well, as expected. For example, the Portuguese *rede* (in English, *network*) is linked to `30-008434259-n`, a synset containing *network* and *web* in the sense of "an interconnected system of things or people", while *ligação* (in English, *connection*) is linked to `30-000145218-n`, a synset containing *joining* and *connection* in the sense of "the act of bringing two things into contact (especially for communication)". However, we also found some domain-specific terms to have been disambiguated incorrectly – for example, the Portuguese *instalação* (in English, *installation*) is linked to `30-003315023-n`, a synset containing *facility* and *installation* in the sense of "a building or place that provides a particular service or is used for a particular industry".

Coreference As mentioned in Sections 9.1.4 and 9.2.4, Batch 1 and 2 of the QTLeap corpus are unusual for coreference, totaling 2000 questions and 2000 corresponding an-

swers. In this context, we also found only a very small number of coreferent pairs (1860 from 82496 markable pairs).

9.4 Summary

For easier reference, Table 24 shows the summary figures for each tool and language, including the three languages that were evaluated in Deliverable 5.4. Note that each figure has been obtained in different evaluation datasets and conditions, and they are not, thus, directly comparable. Still, with the exception of Bulgarian NED, and Basque WSD, the figures seem to indicate that the quality of the processors is in good shape.

Language	NED	WSD	Coreference
Basque	87.90	56.40	53.67
Czech	80.30	80.47	50.28
Portuguese	67.07	65.00	-
Bulgarian	46.88	65.85	50.62
English	77.76	80.10	56.40
Spanish	65.11	79.30	51.38

Table 24: Summary of F1 scores for annotation tools. Top rows for the languages covered in this deliverable. Bottom rows for languages covered in Deliverable 5.4. Note that evaluation sets vary across languages, see text for details.

10 Final remarks

This deliverable has reported the experiments which apply lexical semantics to MT, and more specifically the lexical semantic processing included in Pilot 2. In Pilot 2 we made use of concept resolution, via word sense disambiguation to WordNet, and resolution of domain-specific entities, via gazetteers mined from domain-related resources. The techniques used involve Linked Open Data like WordNet and DBpedia.

The cumulative experiments on Pilot 2 show the improvement of each method. These experiments have been performed over all the languages in WP5 that use the QTLeap TectoMT platform (i.e. Basque, Czech, English, Portuguese, Spanish), except the experiment which tests the contribution of word sense disambiguation, which has been performed on en→pt alone. The main conclusions are the following:

- The best method to incorporate word sense information into translation is to enrich word representations in the MT model, adding word senses as features to the Discriminative TM, as shown by the en→pt results using the QTLeap TectoMT platform. The mere substitution of words by senses did not show improvement.
- Treating domain-specific entities like URLs and commands with the HideIT machinery improves results. It performs better for the translation of answers than questions, as the answers tend to contain such content more frequently than the questions.
- The gazetteers constructed from Wikipedia and other resources are one of the two most beneficial approaches. They are specially useful to translate software texts, e.g., menu items and messages. The gazetteers have been released through META-SHARE.[52]
- The other most beneficial approach is domain adaptation using translation model interpolation, which performed consistently well for both the translations from and to English. This approach smoothly integrates domain specific and general translation models.
- The combination of the lexical semantic techniques mentioned above is highly beneficial for all the languages in both directions. The absolute BLEU improvements in question translation (into English) ranged from 0.77 for in eu→en to 2.50 in cs→en. The improvements in answer translation (from English) was higher, ranging from 1.46 in en→eu to 7.91 in en→es.

In addition, we report additional experiments for lexical semantic techniques which have not been included in Pilot 2. From those experiments, we have learned the following lessons, which we plan to assess in view of its incorporation into Pilot 3:

- The Supersense Tagger WSD software provides sense tags with promising performance gains for en→es using factored Moses for QTLeap queries and News, but not for QTLeap answers. We plan to improve English WSD results combining the Supersense Tagger with UKB, which will hopefully further improve MT performance.

[52] http://metashare.metanet4u.eu/go2/qtleap-specialized-lexicons

- Regarding WSD for answers, the good results using Discriminative TM in TectoMT for en→pt seem to indicate that factored MT has its limitations. We plan to extend the successful technique used in en→pt to the rest of the languages. In addition, we expect that ongoing experiments with efficient machine learning software that allow to build single classifier for all source lemmas will show improvements in the close future. These experiments are checking methods to feed richer word representations like probability distributions for senses and distributional embeddings into the classifier.

- The German experiments also confirmed that WSD helps translating short queries when using factored Moses. In the QTLeap scenario, queries are translated from the other languages into English. We thus plan to make use of the QTLeap WSD systems reported in D5.6 for non-English languages, and check whether we can improve MT performance for translating queries into English.

- The en↔bg experiments, which also use lexical semantics in factored MT, drop the performance with respect to comparable models without lexical semantics. The lack of improvement for en→bg on QTLeap answers agrees with the en→es and en→de results. The lack of improvement for bg→en might be due to the different technique used to encode word sense information. We plan to check whether the techniques used on the successful en→pt, en→es and en→de experiments improve the results of bg→en. Another alternative explanation could be that the current quality of the Bulgarian resources and WSD module is not satisfactory. Note that three of the techniques explored for Bulgarian show improvements over the naive use of word sense information (use of representative target language lemmas, domain-adapted wordnets, and use of coreference information to improve WSD), and we plan to explore their use in the other methods to exploit WSD in MT.

- Detecting named entities with NERC tools and translating using dedicated resources and algorithms in the Moses platform provides very small gains en→es for News texts.

- Building parallel corpora from Wikipedia is effective in the IT domain, as shown for en→es translation using Moses, and we hope to carry this technique to the other language pairs.

- The experiment performed with the integration of coreference resolution shows that there is a role for this procedure, although local rules might suffice in some cases and the use of a full-fledged coreference resolver might not be needed. For instance, resolution of coreference is useful when translating from English to Czech, but less for English to Dutch. It profits mainly from the correct resolution of relative pronouns, the class ignored by many other coreference resolvers.

- We presented an automatic language-independent method for acquiring multiwords from raw text, and integrated these into the TectoMT system. Using source-side analysis only, we demonstrated an improvement in translation quality. By adding target-side analysis during training and generation during testing, we should be able to deliver even larger improvements in future.

As a short outlook on the use of lexical semantic towards Pilot 3, experiments 5.4.2 have finished with promising results for WSD. We think new transduction algorithms are

needed, as planned for the 3rd year in experiment 5.4.5 We also plan to continue examining the contribution of NERC and NED tools in the translation of the News domain, where we expect to have a more prominent role for them.

With regards to the construction of Gazetteers started in Experiment 5.4.3, Experiment 5.4.4 will examine online sources to build gazetteers, which seems a promising direction as shown by the Wikipedia domain corpus experiment.

Finally, this deliverable includes the evaluation of Basque, Czech and Portuguese WSD, NED and Coreference tools, both on standard datasets, and on the QTLeap domain corpora. In general, the figures seem to indicate that the quality of the processors is in good shape, and further improvements will be undertaken. We will seek their use to improve results for language pairs with English as a target language.

References

Itziar Aduriz, Maxux Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Díaz Arantza de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben Urizar. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. *Language and Computers*, 56(1):1–15, 2006.

Rodrigo Agerri, Josu Bermudez, and German Rigau. Ixa pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 26–31, 2014.

Eneko Agirre and Aitor 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, Athens, Greece, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609067.1609070.

Mikel Artetxe, Eneko Agirre, Iñaki Alegria, and Gorka Labaka. Analyzing english-spanish named-entity enhanced machine translation. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 52–54, Denver, Colorado, USA, June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-1007.

Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar Nascimento, Filipe Nunes, and João Silva. Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC '06, pages 1438–1443, 2006.

Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. Prague Dependency Treebank 3.0, 2013.

Nicola Bertoldi and Marcello Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189. Association for Computational Linguistics, 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*. European Language Resources Association, 2012.

Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602, Sydney, Australia, July 2006. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W06/W06-1670.

Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics, July 2002. doi: 10.3115/1118693.1118694. URL http://www.aclweb.org/anthology/W02-1001.

Sandrea Collovini, Thiago I. Carbonel, Juliana Thielsen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. Summit: Um Corpus Anotado com Informações Discursivas Visando à Sumarização Automática. In *Proceedings of the 5th Workshop on Information and Human Language Technology*, TIL'2007, Rio de Janeiro, Brazil, 2007.

Gaoying Cui, Qin Lu, Wenjie Li, and Yirong Chen. Mining concepts from wikipedia for ontology construction. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 287–290. IEEE Computer Society, 2009.

Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

Ondřej Dušek, Eva Fučíková, Jan Hajič, Martin Popel, Jana Šindlerová, and Zdeňka Urešová. Using parallel texts and lexicons for verbal word sense disambiguation. In Eva Hajičová and Joakim Nivre, editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 82–90, Uppsala, Sweden, 2015. Uppsala University, Uppsala University. ISBN 978-91-637-8965-6.

Christine Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Izaskun Fernandez, Iñaki Alegria, and Nerea Ezeiza. Semantic Relatedness for Named Entity Disambiguation Using a Small Wikipedia. In Ivan Habernal and Václav Matoušek, editors, *Text, Speech and Dialogue*, volume 6836 of *Lecture Notes in Computer Science*, pages 276–283. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-23537-5.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. The ontowordnet project: Extension and axiomatization of conceptual relations in wordnet. In Robert Meersman, Zahir Tari, and DouglasC. Schmidt, editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, pages 820–838. Springer Berlin Heidelberg, 2003. ISBN 978-3-540-20498-5. doi: 10.1007/978-3-540-39964-3_52. URL http://dx.doi.org/10.1007/978-3-540-39964-3_52.

Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W08/W08-0509.

Rosa Gaudio and Antonio Branco. Using wikipedia to collect a corpus for automatic definition extraction: comparing english and portuguese languages. In *Anais do XI Encontro de Linguistiva de Corpus - ELC 2012*, Instituto de Ciências Matemáticas e de Computação da USP, em São Carlos/SP, 2012.

Rajdeep Gupta, Santanu Pal, and Sivaji Bandyopadhyay. Improving mt system using extracted parallel fragments of text from comparable corpora. In *proceedings of 6th*

workshop of Building and Using Comparable Corpora (BUCC), ACL, Sofia, Bulgaria, pages 69–76, 2013.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association, 2012.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Comput. Linguist.*, 39(4):885–916, 2013.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195, 2012.

Haibo Li, Jing Zheng, Heng Ji, Qi Li, and Wen Wang. Name-aware Machine Translation. In *ACL 2013*, pages 604–614, 2013.

Lieve Macken, Julia Trushkina, and Lidia Rura. Dutch parallel corpus: Mt corpus and translator's aid. In *In Proceedings of the Machine Translation Summit XI*, pages 313–320. European Association for Machine Translation, 2007.

Paul McNamee and James Mayfield. Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97, 2004. ISSN 1386-4564. doi: 10.1023/B:INRT.0000009441.78971.be. URL http://dx.doi.org/10.1023/B%3AINRT.0000009441.78971.be.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013a.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751, 2013b.

MultiWordNet. The MultiWordNet project. http://multiwordnet.fbk.eu/english/home.php, n.d. Accessed: 2015-01-13.

Steven Neale, João Silva, and António Branco. A Flexible Interface Tool for Manual Word Sense Annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, ISA-11, pages 67–71, London, UK, 2015. Association for Computational Linguistics.

Michal Novák, Dieke Oele, and Gertjan van Noord. Comparison of coreference resolvers for deep syntax translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 17–23, Lisboa, Portugal, 2015. Association for Computational Linguistics.

Hideo Okuma, Hirofumi Yamamoto, and Eiichiro Sumita. Introducing a translation dictionary into phrase-based SMT. *IEICE transactions on information and systems*, 91 (7):2051–2057, 2008.

Santanu Pal, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way. Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*. ACL, 2010.

Magdalena Plamada and Martin Volk. Towards a wikipedia-extracted alpine corpus. In *The 5th Workshop on Building and Using Comparable Corpora*, page 81. Citeseer, 2012.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142, 2011. ISSN 1574-020X.

Martin Popel and Zdeněk Žabokrtský. TectoMT: Modular NLP framework. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233. Springer, 2010.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. Automatic identification of document translations in large multilingual document collections. In *In RANLP 2003 – Proceedings of the International Conference on 'Recent Advances in Natural Language Processing*, pages 401–408, 2003.

Jokin Pérez de Viñaspre. Wikipedia eta anbiguetate lexikala. Technical report, Computer Science Faculty, University of the Basque Country, 2015. URL http://hdl.handle.net/10810/15911.

Uwe Quasthoff and Christian Wolff. The Poisson collocation measure and its applications. In *Proceedings of the 2nd International Workshop on Computational Approaches to Collocations*, Vienna, Austria, 2002.

Bahar Salehi, Paul Cook, and Timothy Baldwin. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 977–983, 2015.

Magda Ševčíková, Zdeněk Žabokrtský, Jana Straková, and Milan Straka. Czech named entity corpus 2.0, 2014. URL http://hdl.handle.net/11858/00-097C-0000-0023-1B22-8. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In *in Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81, 1992.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Díaz de Ilarraza. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. *Procesamiento del Lenguaje Natural*, 55:23–30, 2015. ISSN 1989-7553.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Sophia Ananiadou, and Akiko Aizawa. Normalisation with the BRAT rapid annotation tool. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine*, Zürich, Switzerland, 2012a.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, 2012b. Association for Computational Linguistics.

Andreas Stolcke. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, November 2002.

Jörg Tiedemann. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, 2009.

Zdeňka Urešová. Valenční slovník pražského závislostního korpusu (pdt-vallex). *Studies in Computational and Theoretical Linguistics*, 2011.

Zdenka Uresova, Jana Sindlerova, Eva Fucikova, and Jan Hajic. An analysis of annotation of verb-noun idiomatic combinations in a parallel dependency corpus. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 58–63, Atlanta, 2013. Association for Computational Linguistics.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 9–12. Association for Computational Linguistics, 2008.

Piek Vossen. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography*, 17(2):161–173, 2004. doi: 10.1093/ijl/17.2.161. URL http://ijl.oxfordjournals.org/content/17/2/161.abstract.

Dirk Weißenborn, Leonhard Hennig, Feiyu Xu, and Hans Uszkoreit. Multi-objective optimization for the joint disambiguation of nouns and named entities. In *53nd Annual Meeting of the Association for Computational Linguistics, July*. ACL, 2015.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., 2005.

Deyi Xiong and Min Zhang. A sense-based translation model for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1459–1469, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1137.

Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wikictionary. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008. European Language Resources Association (ELRA).