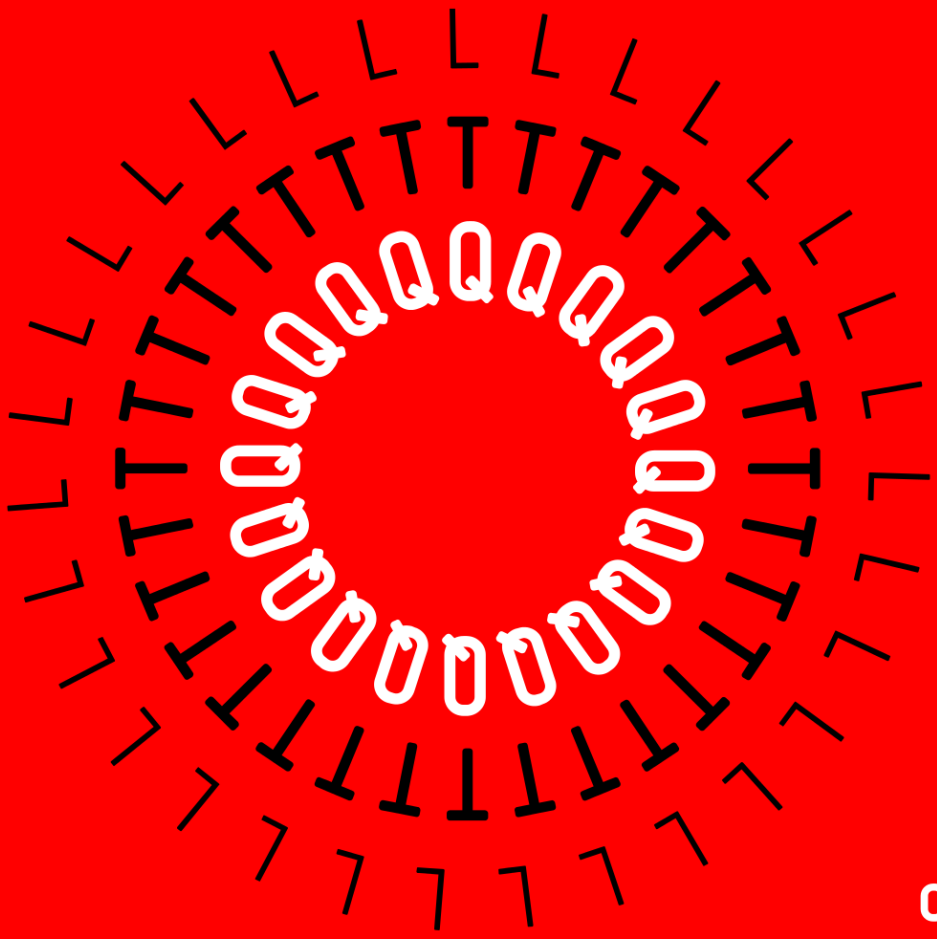


qtleap
quality translation by deep language engineering approaches

Report on the State of the Art Concerning Out-of-vocabulary Expressions

DELIVERABLE D4.2

VERSION 7.0 | 2015-06-04

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:

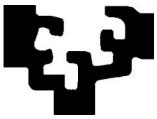
Faculty of Sciences, University of Lisbon

German Research Centre for Artificial Intelligence

Charles University in Prague

Bulgarian Academy of Sciences

Humboldt University of Berlin



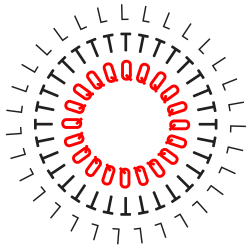
University of Groningen

Higher Functions, Lda

Revision history

Version	Date	Authors	Organisation	Description
1.0	Jan 9, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg	UBER	First draft
1.5	Jan 17, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola	UBER, UPV-EHU	Feedback from UPV-EHU integrated
2.0	Jan 22, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord	UBER, UPV-EHU, UG, IICT-BAS	Feedback from UG and IICT-BAS integrated
3.0	Jan 27, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord	UBER, UPV-EHU, UG, IICT-BAS	Feedback from internal review integrated
3.5	Feb 3, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord, António Branco	UBER, UPV-EHU, UG, IICT-BAS, FCUL	Feedback from FCUL integrated
4.0	Feb 6, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord, António Branco	UBER, UPV-EHU, UG, IICT-BAS, FCUL, DFKI	Minor comments from DFKI addressed
5.0	Apr 28, 2014	Kostadin Cholakov, Valia Kordoni, Markus Egg, Kepa Sarasola, Petya Osenova, Gertjan van Noord, António Branco	UBER, UPV-EHU, UG, IICT-BAS, FCUL, DFKI	Modified in accordance with the recommendations made in the technical audit.
6.0	Oct 8, 2014	Kostadin Cholakov	UBER	Added chapter "Expected Benefits for the Real User Scenario"
6.0	Oct 10, 2014	Kostadin Cholakov	UBER	All relevant partners agreed with the changes made in this version and no further input was provided
7.0	Jun 7, 2015	Kostadin Cholakov	UBER	Changes to address the comments in the first year review

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



REPORT ON THE STATE OF THE ART CONCERNING OUT-OF-VOCABULARY EXPRESSIONS

DOCUMENT QTLEAP-2015-D4.2

EC FP7 PROJECT #610516

DELIVERABLE D4.2

completion

FINAL

status

RESUBMITTED

dissemination level

PUBLIC

responsible

KIRIL SIMOV (WP4 COORDINATOR)

reviewer

ENEKO AGIRRE

contributing partners

FCUL, DFKI, IICT-BAS, UBER, UPV-EHU, UG

authors

KOSTADIN CHOLAKOV, VALIA KORDONI, MARKUS EGG,

KEPA SARASOLA, PETYA OSENOVA, GERTJAN VAN NOORD,

FRANCISCO COSTA, JOÃO SILVA, ALJOSCHA BURCHARDT

Contents

P6

1 Introduction

In the context of the QTLeap project, there are two types of out-of-vocabulary expressions (OOVs), depending on the place they occur in the machine translation (MT) pipeline. The first type are OOVs which are unknown to the native parsing systems of each project language. Such OOVs can cause incorrect parse analyses which will be propagated "up" the MT pipeline and may cause wrong translations. The second type of OOVs are words that do not occur in the parallel data used to train the MT pipeline. As such, no lexical transfer rules can be learnt for those words and they cannot be translated into the target language.

Note that Work Package 4 is about incorporating deep linguistic processing into the MT process. Therefore, the main focus of this deliverable will be on techniques dealing with the first type of OOVs outlined above. However, we will also provide some ideas about handling words not found in the training data for the MT system.

The parsing systems employed in the project are either based on deep linguistic formalisms such as HPSG (Pollard and Sag, 1994) (e.g., Bulgarian, Portuguese) or employ dependencies and other more shallow linguistic information (e.g., German, Spanish). Despite the different formats, most parsing systems use lexicalised grammars in which the lexical information comes from a lexicon or other types of lexical resources. Such systems share a common problem, namely that the lexicon will always be too small. One cannot possibly list every word in a language in the lexicon because natural languages are very productive, constantly changing, hence there will always be new, unseen words.

Improving the lexical coverage of parsing systems is a central issue for their successful integration in real-life applications such as machine translation. If a word is not known to the grammar, parsing stops and, at best only a partial analysis for the input sentence can be provided. In the context of the current project, this issue is even more important because if a word is not in the lexicon of the grammar, it becomes very hard to find a translation for the expression in which this word occurs even if the word is substituted by some "place holder".

A considerable amount of research has been done in devising automated techniques for handling unknown words, especially in deep linguistic processing. As a result, several types of techniques have been developed.

The simplest ones employ heuristics based on orthography and/or n-grams to come up with a proper lexical description whenever the deep grammar encounters an unknown word during parsing (e.g., van Noord and Malouf (2004)). Other, more sophisticated techniques employ statistical classifiers to extend the lexicon offline, i.e. before the grammar is used for parsing (e.g., Cholakov (2012)). A third type of techniques operates online, i.e. during parsing, and employs supertagging (e.g., Clark and Curran (2004)). The main idea behind supertagging is to use a sequential part-of-speech (POS) tagger with a linguistically rich tagset in which the tags encode much more information than simply the POS of the words in the sentence.

Below, we present more details on each of those types of techniques. However, we will first demonstrate the seriousness of the unknown words problem in practice.

2 How Big is the Problem?

Cholakov (2012) performs an interesting experiment in which he illustrates the impossibility of including each word in a language in any lexical resource. He exploits the

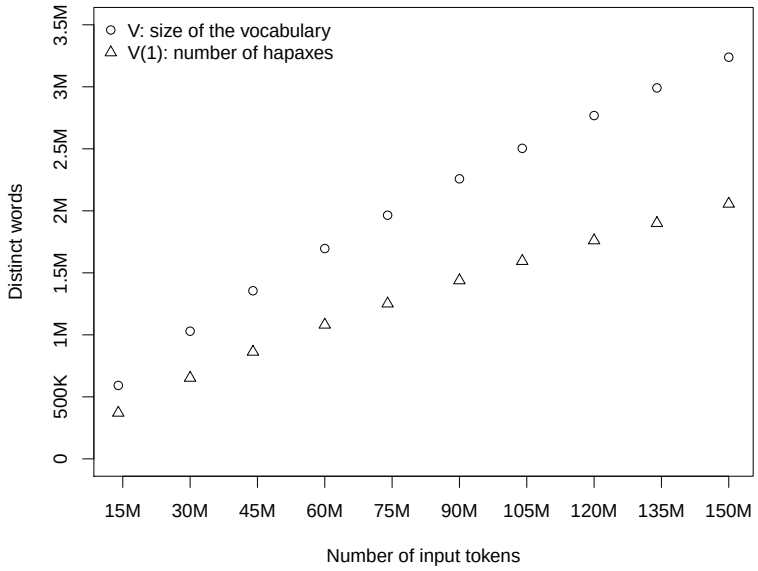


Figure 1: Vocabulary growth and number of hapaxes for a web corpus of 150 million tokens. Note that the number of hapaxes keeps rising despite the increase in the vocabulary size. (The figure is taken from Cholakov (2012).

phenomenon of lexical singletons, i.e. words which occur only once in a given corpus. These singletons are referred to as *hapax legomenon* (plural form: hapax legomena), or hapaxes for short. A web corpus of 150 million word tokens (ignoring punctuation) is lemmatised and POS tagged. Next, word types are counted, i.e. if four different inflected forms of a given word occur in the corpus, those forms will be counted as four occurrences of the same word. Following this manner of counting, it is reported that the corpus contains approximately 3,250,000 distinct words. For comparison, the largest German dictionary, *Das Deutsche Wörterbuch*, contains about 350,000 distinct words. Therefore, a lexicon constructed on the basis of this dictionary will contain only about 11% of the words in the corpus.

Next, suppose that the first 105 million tokens from the corpus are included in some lexicon. If the notion that we can use a very large corpus to include almost all words in a language were applicable in practice, the probability of encountering new words among the remaining 45 million tokens would be extremely low. This means that the number of hapaxes would not increase considerably when the 45 million tokens are processed because we would have already seen almost all words in the language. However, Figure 1 shows that this is clearly not the case. As the number of input tokens increases, the number of hapaxes keeps rising, i.e. there are more and more words encountered which have not been seen among the tokens processed so far.

The same outcome is reported in Fengxiang (2010) for the more balanced British National Corpus (BNC; (Burnard, 2000)) which contains 100 million tokens. Nearly 45% of the distinct words found in the BNC are hapaxes. Further, Fengxiang (2010) implements a computer simulation the result of which shows that, if the size of the corpus approaches infinity, the ratio between the number of hapaxes and the vocabulary size would approach its horizontal asymptote 1, i.e. 50% of all words added to the corpus

would be new. Clearly, no matter how large the corpus we observe is, a lexicon constructed from it cannot contain all words in the language.

Baldwin et al. (2004) illustrate the problem unknown words pose to deep parsing specifically. The work examines the performance of the large-scale HPSG English Resource Grammar (ERG; Copestake and Flickinger (2000)) on a sample of 20,000 sentences which have been randomly extracted from the BNC corpus. The experiments show that 68% of the sentences contain at least one unknown word. Nouns account for the largest portion of unknown words (61%), followed by adjectives (22%) and verbs (13%). Similar results are presented in Zhang and Kordoni (2006) where a newer version of the ERG was employed.

Cholakov et al. (2008) report that 76% of the sentences in a German newspaper corpus (ca 615,000 sentences) contain words which were not listed in the lexicon of the large-scale GG grammar (Crysmann, 2003) of German. When the grammar is applied to 2.57 million sentences extracted from web corpora, this percentage increases to 90.5%.

Thomforde and Steedman (2011) observe that a CCG parser trained on sections 2-21 of the deep English CCGBank treebank (Hockenmaier and Steedman, 2007) and tested on Section 00 of this treebank encounters 20% unknown tokens.

Next, we present an overview of the techniques which have been used to deal with unknown words in the context of deep linguistic processing.

3 Offline Techniques

Offline techniques aim at extending the lexicon of the grammar with all possible categories of a given unknown word. The strategy behind those techniques is to improve the parsing coverage and accuracy by improving the quality and the coverage of the lexicon of the grammar before the actual parsing is performed. This means that such techniques are independent from the parsing input and thus concentrate on a more permanent solution to the problem of unknown words. Online methods, on the other hand, acquire only the particular lexical category which an unknown word has in a given input sentence.

The most prominent offline technique is presented in Baldwin (2005). It is applied to the task of predicting the lexical categories for a test set of words which are temporarily removed from the lexicon of the ERG, thus making those unknown to the grammar. The technique involves several language resources: character n-grams derived from the unknown words, a POS tagger, a dependency parser, a chunker, and an ontology (WordNet 2.0 (Fellbaum, 1998)). Those resources are used to process sentences in which the unknown words occur and then various features for classification are extracted from the output produced by each resource. This means that the prediction process is based on information which comes outside the deep grammar. The advantage of employing various language resources lies in the possibility that those resources may provide linguistic information which is absent from the grammar. This makes the technique of Baldwin (2005) particularly useful for resource light languages with less evolved deep grammars which do not yet contain a sufficient amount of linguistic information.

Except for the experiments with WordNet, a set of 110 binary classifiers is constructed, one for each target lexical category in the ERG lexicon. A positive outcome of the classifier for a given unknown word means that the word belongs to the lexical category this classifier was trained for. Therefore, a word can be assigned multiple lexical categories. A potential shortcoming of this architecture is that a word can be negatively classified by all classifiers and thus, no lexical category will be assigned to it.

The involvement of WordNet is motivated by the hypothesis that there is a strong correlation between syntactic and semantic similarity of words, as best exemplified in the work of Levin (1993). The set of binary classifiers is not applied to WordNet. Instead, the lexico-semantic hierarchy adopted in WordNet is used to extract all synonyms, direct hypernyms and direct hyponyms for a given test word. Then, a majority vote is taken across the lexical categories of those extracted semantic "neighbours".

The model involving the POS tagger proves to be the most successful. It is, however, the only model the performance of which is above that of a baseline model assigning to a given test word the most frequently occurring lexical category for the POS of this word in the ERG lexicon.

Another offline technique is presented in Savkov et al. (2011) who employ an external POS tagger based on Support Vector Machines (SVM) and a rule-based computational morphology in addition to the internal guessing techniques of the grammar. If the outputs of the three techniques for a given unknown word disagree, all three lexical descriptions are kept. If the morphology and the guesser produce identical descriptions, the output of the POS tagger is discarded. Finally, if the POS tagger is the only technique which produces a description for the unknown word, this description is adopted by the grammar.

In conclusion, while offline techniques may seem a promising way to handle unknown words because those concentrate on permanent improvements to the lexicon of the deep grammar, those methods suffer from the very same problem they are trying to solve. Even if a large good-quality extension of the lexicon can be done, the grammar will still suffer from insufficient lexical coverage.

Furthermore, offline techniques are not suited for the purposes of QTLeap which aims at integrating robust deep linguistic processing with MT. While such techniques can improve the performance of the grammar for some fixed corpora, they will always be pruned to the constant lexical creativity of the users of real-life MT systems. Finally, offline techniques are generally very sophisticated and time-consuming which makes it nearly impossible to convert them into real-time applications.

4 Online Techniques: Ad hoc Techniques

In order to integrate deep linguistic processing with MT, real-time online techniques for handling unknown words are needed. The least sophisticated techniques include ad hoc unknown word guessers. Those guessers rely on morphology, capitalisation, numbers, and other orthographic properties of the unknown word. Such techniques typically try to provide some general information about the unknown words (e.g., parts-of-speech) which does not represent a full lexical description as defined in the grammar. The sole purpose of these partial descriptions is to enable the grammar to produce some parse for a sentence containing unknown words even if this parse contains less information than that of a sentence without any unknown words. An example of such a technique is the unknown word guesser presented in van Noord and Malouf (2004) which is part of the Alpino parser and grammar of Dutch (van Noord, 2006).

The NLP group at the University of the Basque Country has developed a guesser for Basque based on morphosyntactic information using finite-state technology (Ezeiza et al., 1998). The set of possible tags is based on open lexical categories only. However, different tagsets of different linguistic granularity can be used in the disambiguation process. A four-level framework is defined guided by pragmatical reasons: 20 tags belong in the first level, 48 in the second, close to 300 in the third, and a couple of thousands in the fourth

level. Overall, the average ambiguity rate in text ranges from 1.5 at the first level to 3.5 at the fourth level. P11

The process of disambiguation is then carried out by combining Hidden Markov Models (HMMs) and rule-based methods. The tagger provides full morphosyntactic information. The third level has been evaluated as the level providing the best balance between fine-grained linguistic information and precision for disambiguation.

Similar to the employment of HMMs, other methods also make use of the context of the unknown word to obtain clues for its proper lexical category. In Fouvry (2003), the grammar is directly used to generate lexical entries which contain fewer or no constraints for all unknown words encountered during parsing. The assignment of fewer constraints to a given word leads to a larger combinatorial potential for that word. This makes it easier for the parser to produce full parses for a sentence containing unknown words. Then, by collecting the necessary information from these full parses, new constraints can be added to the lexical entries assigned to the unknown words. These constraints specify the proper linguistic information for those words.

Thomforde and Steedman (2011) explore a similar idea of using the grammar and the context of the unknown word. The partial parse chart formed from analysing the known words surrounding a given unknown word and the categories of these words are used to infer the category of the unknown word. The result of the process is a ranked set of possible lexical categories for this word. This technique is referred to as chart mining.

A common problem with these ad hoc techniques is that, generally, they do not provide the full amount of linguistic information which a lexical entry in the lexicon provides. Further, using underspecified lexical entries or chart mining increases the combinatorial potential of unknown words and leads to a higher number of possible analyses for a given sentence. This increases ambiguity and makes it harder for the parser to choose a correct analysis. Finally, techniques which rely on word context become less reliable if two or more unknown words are adjacent.

5 Supertagging

5.1 Main Idea

Another online technique which may be more suitable for the purposes of QTLeap is supertagging. Generally, supertagging refers to the process of applying a sequential tagger to assign lexical descriptions associated with each word in an input string, relative to a given grammar. In comparison to a "standard" POS tagger, a supertagger has a larger and a more linguistically refined tagset. The tags in this tagset are either the lexical categories found in the grammar or some abstract form thereof. In both cases, the tags are usually more informative than plain POS tags.

Supertagging was introduced by Bangalore and Joshi (1999) as a technique for reducing parsing ambiguity of lexicalised tree adjoining grammars grammars (LTAGs), and has since been applied within CCG (Clark, 2002; Clark and Curran, 2004) and HPSG (Dridan et al., 2008; Dridan, 2013; Zhang et al., 2010b) as well as with the Dutch Alpino grammar (Prins and van Noord, 2003). Since the tagger assigns lexical descriptions to each word in the input, a side effect of supertagging is the assignment of descriptions to unknown words as well. That is why supertagging has also been employed specifically for learning unknown words. Note that in this case the evaluation of the supertagger is performed only for the lexical categories assigned to the unknown words in the input. In the "standard"

application of supertagging, the tagger is evaluated on the whole input. P12

The supertagger can be trained with various types of features. Most approaches concentrate on lexical features and some simple features extracted from the context of the unknown word (e.g., surrounding words). Some more recent approaches have also integrated structural features extracted from syntactic analyses. Below, we present both types of methods.

5.2 Supertagging Using Deep Lexical Categories

The first significant work which explicitly applies supertagging to the task of learning unknown words is that of Zhang and Kordoni (2006). This work has initiated a new line of research and it has inspired the development of multiple similar techniques for deep grammars of English, German, and Dutch.

The suppertagger comes in the form of a dedicated maximum entropy-based (ME-based) classifier. It is applied to the task of predicting lexical categories for words unknown to the ERG grammar. Normally, there are great many lexical categories in the lexicons of large-scale deep grammars such as the ERG. Therefore, the tagger should be able to handle a large number of possible outputs as well as handle thousands of features. Those factors are the main motivation for the choosing ME as the classification algorithm. ME-based classifiers have also the advantage of general feature representation and no independence assumption between features.

The tagger is trained on the Redwoods treebank (Oepen et al., 2004) which then consisted of approximately 11,000 sentences, parsed with the ERG and corrected manually. A 10-fold cross validation is performed on these data. The words occurring in the test fold are treated as unknown. Since each word in the treebank is annotated with the correct lexical type, it is straightforward to construct a gold standard for evaluation.

The results of the tagger are compared to a baseline which assigns the majority lexical category in the lexicon for the POS of the unknown word. The performance of the supertagger is also compared to the state of the art TnT POS tagger (Brants, 2000) which is a very efficient POS tagger based on HMMs. In this setup, the tagset of TnT contains lexical categories from the ERG lexicon instead of simple POS tags. The experiments show that the supertagger achieves the best performance in all experimental scenarios, achieving about 60% accuracy.

Another novelty in the approach of Zhang and Kordoni (2006) is the addition of a post processing step which further improves the performance of the supertagger. After the tagger outputs a ranked list of candidate lexical categories for a given unknown word, the top 3 categories are assigned to this word. Then, the sentence containing the word is parsed again and the parser selects the best analysis. The lexical category assigned to the unknown word in this analysis is taken as the final outcome. Such an approach has the benefit of allowing the grammar itself to choose the category which it "considers" to be best suited for the word in that particular context. This means that, for learning unknown words, we can exploit to the fullest extent the linguistic information encoded in the grammar.

The same supertagger but without the post processing step is used in Nicholson et al. (2008) with the GG grammar of German. The purpose is to examine the application of the technique to other languages which also have richer morphology than English and thus, may pose a bigger challenge for the supertagger. A similar experiment with the GG is presented in Cholakov et al. (2008), which we will discuss in more detail later on.

We should also mention two other techniques based on standard POS taggers, which are slightly modified to handle deep grammar lexical categories. Baldwin (2005) evaluates the application of a supertagger to the task of learning lexical categories for words unknown to the ERG. The supertagger employed is an out-of-the-box trainable POS tagger, in the form of fnTBL 1.0 (Ngai and Florian, 2001). fnTBL is a transformation-based learner that is distributed with pre-optimised POS tagging modules for English and other European languages. In the experiments presented in Baldwin (2005), however, the tagset of the tagger consists of 110 open-class lexical categories identified in the lexicon of the ERG: noun, adjective, verb and adverb types. The tagger is trained on the Redwoods treebank.

A further development of this technique is presented in Blunsom and Baldwin (2006). The supertagger employed there is based on pseudo-likelihood conditional random fields (CRF). CRF support the use of a large number of non-independent and overlapping features of the input. This property of CRF allows for the inclusion of additional features to be used during tagging. In particular, Blunsom and Baldwin (2006) employ various orthographic features which include prefix and suffix information and binary features which encode information about whether the unknown word contains characters from some pre-defined character set. The tagger is applied to predict lexical categories for words unknown to the ERG and Jacy, a large HPSG grammar of Japanese (Siegel and Bender, 2002).

5.3 Supertagging Involving Syntactic Information

Research efforts have also been made to incorporate syntactic information in supertagging. The assumption behind is that features describing inter-word dependencies will provide for a more accurate assignment of lexical categories to unknown words. This is especially important for verbs which can have multiple types of subcategorisation frames.

Matsuzaki et al. (2007) proposes a supertagger for the Enju HPSG grammar of English in which a context-free-grammar (CFG) is used to filter the tag sequences produced by the tagger before running the parser. In this case, the CFG is an approximation of the HPSG grammar. The key property of this CFG approximation is that the language it recognises is a superset of the parsable tag sequences produced by the supertagger. Hence, if the CFG fails to parse a sentence, the respective tag sequence is discarded, thus reducing the amount of tag sequences the parser has to consider.

Despite the fact that the primary purpose of this technique is to reduce parsing times, the quality of the supertagging can be inferred from the increase in parsing accuracy when the supertagger is employed. This increase is partly due to the improved prediction of lexical categories for unknown words compared to the default guessing techniques used when the supertagger is not employed.

Silva and Branco (2012a) and Silva and Branco (2012b) employ a supertagger based on support-vector machines with the LXGram HPSG grammar of Portuguese (Branco and Costa, 2008). Training data for the tagger is created semi-automatically. The grammar is used to parse a corpus and human annotators then choose the correct analysis from all analyses proposed by the grammar for a given sentence.

First, dependency relations are extracted from the parses in the training data. Those relations are represented as a list of tuples, each of which relates a pair of words in a given sentence through a grammatical relation. Further, each word is also annotated with its POS tag, lexical category (as defined in the lexicon of the grammar) and lemma.

Then, features are extracted for training the supertagger. Since the full dependency representation of a sentence contains a lot of features, irrelevant to the task of predicting lexical categories for unknown words, only those words are considered which are directly connected to the target word via a grammatical relation.

Another interesting aspect of this approach is the usage of SVM for learning unknown words. SVM are binary classifiers while unknown word prediction for deep grammars is a multi-class classification problem. The authors had to binarize the problem first by creating a classifier for each possible pair of lexical categories. Despite this additional effort, using SVM pays off. In experiments in which the supertagger was used to handle unknown words in Portuguese newspaper and Wikipedia texts, it outperformed the HMM-based TnT POS tagger. The tagsets for TnT and the supertagger were the same.

5.4 Combining Online and Offline Techniques

Research in unknown words learning clearly shows that it is important to incorporate the linguistic knowledge contained in the grammar (Zhang and Kordoni, 2006) but it is as equally important to provide additional information as well as consider various contexts in which the unknown word occurs (Baldwin, 2005). Cholakov (2012) combines the advantages of those research ideas. He presents a statistical technique which involves the grammar directly into the learning process for unknown words. Furthermore, the lexical categories for the whole paradigm of a given unknown word are learnt while other approaches are only concerned with particular word forms. The core of the technique is a maximum entropy based classifier which takes morphological and syntactic features as input and outputs lexical categories. Thus two factors are considered – the morphology of the unknown word and the syntactic constraints imposed by its context.

As for the former, the acquisition of the whole paradigm provides a valuable source of morphological information. If only one form of the unknown word were taken into account, this information would not be accessible. Morphological features include character n-grams, orthographic properties, and the morphological type of the unknown word as predicted by an external finite state morphological analyser.

Syntactic features, on the other hand, are extracted from parsing a large number of sentences containing a given unknown word with a deep parser. For this purpose, each unknown word is initially assigned all target lexical categories, thus letting the parser itself choose a proper category for each input sentence. The categories assigned by the parser are then used as features in the statistical classifier. Finally, looking at different contexts of the unknown word provides the possibility to work with linguistically diverse data and to incorporate more syntactic information into the learning process. Cases where this is particularly important include morphologically ambiguous words and verbs which subcategorise for various types of syntactic arguments. Contexts of the other members of the paradigm of the unknown word are also considered in order to increase the amount of linguistic data the method has access to.

This technique is considered largely language- and formalism-independent. This point was demonstrated by applying it successfully to the Dutch Alpino grammar and the GG grammar of German. In both studies the achieved parsing coverage and accuracy were better compared to the setups in which other techniques for dealing with unknown words had been used. Although the technique is not entirely an online one, its basic ideas and research contributions could be adapted to an online application scenario.

5.5 Granularity Matters

Finally, we should underline an important aspect of supertagging, namely the granularity of the tagset, i.e. the amount of linguistic information encoded in the tags. More detailed and refined information means a larger number of tags which, in turn, makes it more difficult for the supertagger to choose the correct sequence of tags. On the other hand, fewer make the task easier but such tags also contain less information. This might lead to worse parsing performance because the lexical description created for an unknown word might be too general, thus allowing for a huge number of combinations for this word and increasing parsing ambiguity. Therefore, one must find the right balance between tag expressiveness and tag predictability.

For example, Cholakov et al. (2008) use the same supertagger setup as Nicholson et al. (2008) to handle unknown words in experiments with the GG grammar. The main difference is that the tagset for the tagger consists of much more linguistically rich tags. The tagger experiences a small decrease in unknown word prediction accuracy but when incorporated into the GG setup, parsing accuracy increased compared to that reported in Nicholson et al. (2008). The authors conclude that despite the underperformance of their method, the tags assigned to the unknown words provide more detailed linguistic information which, in turn, improves parsing accuracy.

Dridan (2009) performs experiments with the ERG and different supertaggers, including the TnT tagger, as well as tagsets of various granularity. TnT shows the best performance when the tagset consists of only 13 POS tags (97%accuracy). However, in another, very detailed setup, in which each tag is formed by the lexical category concatenated with any selectional restriction in the lexical entry (803 tags), the accuracy of TnT decreases to 91%.

In conclusion, supertagging has made considerable improvements in the handling of unknown words online, i.e. during parsing. It provides a good basis for the research we plan to do within the QTLeap project with regard to unknown words. The next section outlines the main research directions we will pursue to improve on state of the art techniques.

6 Beyond the State of the Art

The high-quality machine translation systems of the future require techniques for robust semantic processing. Statistical MT based on word or phrase alignment and the employment of large (parallel) corpora has led to a significant improvement of translation quality. However, such techniques are already reaching their limits. The next step is to employ more sophisticated methods which provide for a deeper natural language understanding. In a nutshell, we need *semantics*.

In the context of QTLeap, where MT for technical texts is performed, semantics will indicate that the word *menu*, for example, should be translated into the technical sense rather than into the gastronomic one. That is why the project seeks the incorporation of robust deep grammars into MT because those grammars can provide detailed semantic representations. In MT, robustness is a key issue and from what we presented above, it is clear that unknown words are the main obstacle for the robustness of deep grammars. Therefore, an imminent research goal in this project is the development of more robust learning techniques for unknown words. The most immediate benefit of such techniques will be the propagation of high quality parsing and semantic analysis of OOVs up the MT

pipeline. This will lead to better semantic transfer thus to more correct translations. P16

To this extent, we need to go further and develop learning techniques which can process unknown words *semantically*. Generally, today's methods for handling unknown words are concerned with the acquisition of the morphological and syntactic properties of unknown words. While this can improve the syntactic coverage and accuracy of the grammar, it is unknown whether the semantic representation (if any) produced by the grammar for a given unknown word in the final sentence analysis is correct. It might be the case that the lexical category acquired for some unknown word enables a correct syntactic parse with a wrong semantic representation. It is therefore vital to also concentrate research efforts on robust *semantic* acquisition for unknown words, thus assuring that the grammar is producing a correct semantic output.

Our main goal is to employ existing techniques for unknown word prediction which already provide good morphosyntactic analyses with robust semantic processing, so that a (partial) semantic representation is generated for each input sentence. Good quality prediction of the morphosyntactic properties of unknown words will allow us to construct proper semantic representations. A good starting point for the research in this project is the work presented in Zhang et al. (2010a) on lexical acquisition via parse chart mining. Extending this work to include semantic features will allow for accurate semantic descriptions to be produced for unknown words.

Another line of research will build on the work initiated in Silva and Branco (2012a,b) and further explore the use of structured features that are obtained from deep treebanks in supertagging. However, those features will also encode semantic information extracted from the semantic representations in the respective deep treebank. Those representations might come in various forms: dependency graphs, MRS, etc. This will allow the supertagger to learn the semantic properties of unknown words based on the semantic information provided by their context as well as based on the types of semantic relations those words participate in.

The two outlined methods are not mutually exclusive but rather complete one another and represent a good basis for further research on learning the semantic properties of unknown words. The semantic representations produced could be in the form of MRS for the German, Bulgarian, Portuguese, and English deep grammars. The tectogrammatical format of the Prague Dependency Treebank describes semantic roles rather than grammatical functions. However, this format can also be mapped to MRS representations, as demonstrated in Jacob et al. (2010).

Consistent with work package 5, we will also incorporate information from Linked Open Data when constructing semantic representations for unknown words. The unknown word can be linked to some lexical semantic ontology or database such as WordNet or VerbNet (Kipper et al., 2008) as well as to articles in encyclopaedic online resources such as Wikipedia. Also, note that many unknown words represent named entities which can easily be linked to entries in a semantic database or Wikipedia. In practice, this means a linking between the grammar dictionary and a particular entry in a given LOD resource (via word sense disambiguation). This way, the knowledge contained in the grammar can be enriched with information from LOD resources which, in turn, will lead to the production of more detailed MRS structures. This will facilitate an MT transfer based on deep linguistic processing.

Another possibility for obtaining semantic information is the usage of distributional models, for example word embeddings. We will build distributional models for all languages targeted in QTLeap and whenever the native system for a given language encoun-

ters an unknown word, it may consult those models for closely related words which are known. This way the system could guess the linguistic properties of the unknown word based on those of closely related known words.

Finally, within the project, we will improve the handling of unknown words based on the data which the other partners in the project will provide. Our initial research will most likely concentrate on English, German, and Bulgarian. In close cooperation with the relevant partners, we will also transform the obtained parsing output into a format which is suitable for use in the MT prototypes and systems developed in QTLeap. The report on a first pilot of enhanced deep language processing systems will be reported by the end of month 12 of the project, as stated in the Description of Work of QTLeap. A second report will be prepared by the end of month 24. Those reports will be prepared in close cooperation with the other partners in Work Package 4 since the reports are concerned with improving the robustness of deep processing systems as a whole. This includes improved treatment of unknown words as well as improved semantic processing and natural language generation.

7 Expected Benefits for the Real Usage Scenario

The work done within the project on improving the quality of MT will be showcased in a real-life usage scenario. In this scenario, clients make requests or pose questions to an IT-helpdesk. These data are provided by the project partner HF and include short sentences, usually a request for help followed by an answer. In such a QA scenario the questions and answers are quite short, but even in such cases deep semantics would be useful, since the failure to analyse and translate the data correctly would lead to total miscommunication.

7.1 OOVs in Parsing

Deliverable D4.1. provides clear examples in support of this claim. This deliverable also shows that great many problems such as wrong lexical choices, incorrectly translated long distance dependencies, etc can be solved if correct semantic interpretations in the form of MRS or some other deep semantic formalism are employed. As mentioned above, lexicons play crucial role in deep linguistic processing. It is vital for the grammar to have proper and detailed lexical entries for all words in a sentence in order to parse it correctly and build correct semantic interpretations which the MT system can in turn benefit from.

An experiment with the ERG deep grammar of English and a subset of the HF data showed that about 87% of the sentences contained a word which was not listed in the lexicon of the grammar. This shows that unknown words pose a huge problem despite the fact that the sentences in the data are relatively short and simple. The unknown words encountered included not only technical terms (which is not a surprise given the strict domain orientation of the data) but also a significant number of every-day words. Another challenge is posed by words which are in the lexicon of the grammar but are used in different contexts and have different linguistic properties within the technical domain which the project data is derived from.

Note that the ERG is a large-scale high quality grammar. Deep grammars for other languages in the project tend to be much smaller and thus the issue of unknown words will be even more severe for them. This illustrates clearly the necessity of dedicated techniques

for handling unknown words and for assuring that correct semantic interpretations are produced for those words. P18

7.2 OOVs in the MT System

As mentioned in the introduction, there is a second type of OOVs, namely words that do not occur in the parallel data used to train the MT system. This is a serious issue in QTLeap because the data which are translated come from technical domains for which there is insufficient amount of parallel data. Experiments in Pilots 0 and 1 show that such OOVs have a huge impact on the results, especially for Basque and Spanish. In those languages even simple technical terms such as *password* cannot be translated due to lack of suitable parallel data.

There are a couple of possibilities to deal with this type of OOVs within the work done in WP 4. One possibility is to build lexicons from software platforms which are provided in multiple languages. For example, there are freely available translations of all menus and help content for LibreOffice[1] and VLC Media Player.[2] Such lexicons can be used as a fall-back strategy whenever the MT model does not contain a transfer rule for a given word.

Another possibility is to consult multilingual ontologies like the ones being constructed in WP 5 or an online resource such as Wiktionary.[3] Experiments with Spanish showed that such resources contain translations for a large amount of OOVs in the QTLeap data. Naturally, some words are polysemous and thus have multiple translations. In such cases, disambiguation heuristics will be employed to pick the correct translation. Once more, this work will be linked to disambiguation activities being performed in WP 5.

8 Conclusion

This deliverable outlined the need for robust deep linguistic processing in the context of MT and identified words not listed in the lexicons of deep grammars as the main obstacle on the way to achieving this robustness. We demonstrated that is not possible for grammars to include each word in a language in their lexicons. In fact, experiments showed that, in real-life scenarios, there would always be a large number of unknown words.

We then described various techniques of different complexity for handling unknown words. Some of those are not suited for the QTLeap project since they cannot be applied in real time, i.e. while parsing. Clearly, an MT system needs an online solution.

The discussed online techniques, however, are not of sufficient quality in order to be applied directly in the current project. The most serious issue is the inability of those methods to learn the semantic properties of unknown words. Semantics is of vital importance to modern MT and therefore, we stated that, within QTLeap, we would pursue robust semantic processing. This will be achieved by enriching existing techniques for unknown word handling (supertagging, chart mining) with semantic features, by incorporating information from LOD resources, and by employing distributional similarity techniques.

[1] https://www.libreoffice.org/
[2] http://www.videolan.org/vlc/index.html
[3] http://www.wiktionary.org/

The deliverable also touched upon OOVs with regard to MT, i.e. words that are not found in the parallel data used for training the MT system. For this type of OOVs, we will closely cooperate with WP 5 to construct large translation lexicon based on multilingual software manuals, multilingual ontologies, and online resources such as Wiktionary. Those lexicons can then be used as a fall-back strategy by the MT system. P19

Bibliography

Baldwin, T. (2005). General-purpose lexical acquisition: Procedures, questions and results. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 23–32, Tokyo, Japan.

Baldwin, T., Bender, E., Flickinger, D., Kim, A., and Oepen, S. (2004). Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2047–2050, Lisbon, Portugal.

Bangalore, S. and Joshi, A. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–65.

Blunsom, P. and Baldwin, T. (2006). Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of EMNLP 2006*, pages 164–171, Sydney, Australia.

Branco, A. and Costa, F. (2008). A computational grammar for deep linguistic processing of Portuguese: LX-Gram. Technical report, University of Lisbon.

Brants, T. (2000). TnT – A statistical part-of-speech tagger. In *Proceedings of the 6th Appied Natural Language Processing Conference (ANLP)*, Seattle, WA.

Burnard, L. (2000). *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services Oxford.

Cholakov, K. (2012). *Lexical Acquisition for Computational Grammars: A Unified Model*. PhD thesis, University of Groningen, The Netherlands.

Cholakov, K., Kordoni, V., and Zhang, Y. (2008). Towards domain-independent deep linguistic processing: Ensuring portability and re-usability of lexicalised grammars. In *Proceedings of COLING 2008 Workshop on Grammar Engineering Across Frameworks (GEAF08)*, pages 57–64, Manchester, UK.

Clark, S. (2002). Supertagging for combinatory categorical grammar. In *Proceedings of the 6th International Workshop on Tree Adjoining Grammar and Related Frameworks*, pages 101–106, Venice, Italy.

Clark, S. and Curran, J. (2004). The importance of supertagging for wide-coverage CCG parsing. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'2004)*, pages 282–288, Geneva, Switzerland.

Copestake, A. and Flickinger, D. (2000). An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resource and Evaluation (LREC 2000)*, Athens, Greece.

Crysmann, B. (2003). On the efficient implementation of German verb placement in HPSG. In *Proceedings of RANLP 2003*, Borovets, Bulgaria.

Dridan, R. (2009). *Using Lexical Statistics to Improve HPSG Parsing*. PhD thesis, Saarland University, Germany.

Dridan, R. (2013). Ubertagging. Joint segmentation and supertagging for English. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1201–1212, Seattle, USA.

Dridan, R., Kordoni, V., and Nicholson, J. (2008). Enhancing performance of lexicalised grammars. In *Proceedings of ACL-08: HLT*, pages 613–621, Columbus, Ohio.

Ezeiza, N., Aduriz, I., Alegria, I., Arriola, J. M., and Urizar, R. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of COLING-ACL 1998*, pages 380–384, Montreal, Canada.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Fengxiang, F. (2010). An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics*, 36(4):631–637.

Fouvry, F. (2003). Lexicon acquisition with a large-coverage unification-based grammar. In *Companion to the 10th Conference of EACL*, pages 87–90, Budapest, Hungary.

Hockenmaier, J. and Steedman, M. (2007). CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Jacob, M., Lopatková, M., and Kordoni, V. (2010). Mapping between dependency structures and compositional semantic representations. In *Proceedings of LREC 2010*, pages 2491–2497.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, 42:21–40.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Matsuzaki, T., Miyao, Y., and Tsujii, J. (2007). Efficient HPSG parsing with supertagging and CFG-fitering. In *Proceedings of the 20th Joint Conference on Artificial Intelligence*, pages 1671–1676.

Ngai, G. and Florian, R. (2001). Transformation-based learning in the fast lane. In *Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8, Pittsburgh, Pennsylvania.

Nicholson, J., Kordoni, V., Zhang, Y., Baldwin, T., and Dridan, R. (2008). Evaluating and extending the coverage of HPSG grammars: A case study for German. In *Proceedings of LREC-2008*, pages 3134–3137, Marrakesh, Morocco.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. (2004). Lingo Redwoods. *Research on Language & Computation*, 2(4):575–596.

Pollard, C. and Sag, I. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Prins, R. and van Noord, G. (2003). Reinforcing parser preferences through tagging. *Traitment Automatique des Langues*, 44:121–139.

Savkov, A., Laskova, L., Osenova, P., Simov, K., and Kancheva, S. (2011). A web-based morphological tagger for Bulgarian. In *Proceedings of the Sixth International Conference on Natural Language Processing, Multilinguality*, pages 126–137, Bratislava, Slovakia.

Siegel, M. and Bender, E. (2002). Efficient deep processing of Japanese. In *Proceedings of the 3rd workshop on Asian language resources and international standardization*, pages 1–8, Taipei, Taiwan.

Silva, J. and Branco, A. (2012a). Assigning deep lexical types. In *Proceedings of TSD2012 – The 15th International Conference on Text, Speech and Dialogue*, pages 240–247, Brno, Czech Republic.

Silva, J. and Branco, A. (2012b). Assigning deep lexical types using structured classifier features for grammatical dependencies. In *Proceedings of the Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages at the 50th Annual Meeting of the Association for Computational Linguistics*.

Steedman, M. (2001). *The syntactic process*. The MIT press.

Thomforde, E. and Steedman, M. (2011). Semi-supervised CCG lexicon extension. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1246–1256, Edinburgh, UK.

van Noord, G. (2006). At last parsing is now operational. In *Proceedings of TALN*, pages 20–42, Leuven, Belgium.

van Noord, G. and Malouf, R. (2004). Wide coverage parsing with stochastic attribute value grammars. In *Draft available from http://www. let. rug. nl/vannoord. A preliminary version of this paper was published in the Proceedings of the IJCNLP workshop Beyond Shallow Analyses, Hainan China*, volume 2005.

Zhang, Y., Baldwin, T., Kordoni, V., and Martinez, D. (2010a). Chart mining-based lexical acquisition with precision grammars. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, Los Angeles, CA, USA.

Zhang, Y. and Kordoni, V. (2006). Automated deep lexical acquisition for robust open text processing. In *Proceedings of the 5th International Conference on Language Resourses and Evaluation (LREC 2006)*, pages 275–280, Genoa, Italy.

Zhang, Y., Matsuzaki, T., and Tsujii, J. (2010b). A simple approach for HPSG supertagging using dependency information. In *Proceedings of 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'10)*, pages 645–648, Los Angeles, CA.