D1.4: Scientific Report on Rich Tree-Based SMT

Ondřej Bojar, Mauro Cettolo, Silvie Cinková, Philipp Koehn, Miroslav Týnovský, Zdeněk Žabokrtský

Distribution: Public

EuroMatrixPlus
Bringing Machine Translation
for European Languages to the User

ICT 231720 Deliverable D1.4

Revision 324, 2010-03-14 23:16:53 +0100 (Sun, 14 Mar 2010)

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development.

Project ref no.	ICT-231720
Project acronym	EuroMatrixPlus
Project full title	Bringing Machine Translation for European Languages to the User
Instrument	STREP
Thematic Priority	ICT-2007.2.2 Cognitive systems, interaction, robotics
Start date / duration	01 March 2009 / 38 Months

Distribution	Public
Contractual date of delivery	February 28, 2012
Actual date of delivery	March 31, 2012
Deliverable number	D1.4
Deliverable title	Scientific Report on Rich Tree-Based SMT
Type	Report
Status & version	Final, (revision 324)
Number of pages	16
Contributing WP(s)	WP1
WP / Task responsible	CU
Other contributors	UEDIN, FBK
Internal reviewer	Chris Callison-Burch
Author(s)	Ondřej Bojar, Mauro Cettolo, Silvie Cinková, Philipp Koehn, Miroslav Týnovský, Zdeněk Žabokrtský
EC project officer	Michel Brochard

The partners in EuroMatrixPlus are:	DFKI GmbH, Saarbrücken (DFKI)
	University of Edinburgh (UEDIN)
	Charles University (CUNI-MFF)
	Johns Hopkins University (JHU)
	Fondazione Bruno Kessler (FBK)
	Université du Maine, Le Mans (LeMans)
	Dublin City University (DCU)
	Lucy Software and Services GmbH (Lucy)
	Central and Eastern European Translation, Prague (CEET)
	Ludovit Stur Institute of Linguistics, Slovak Academy of Sciences (LSIL)
	Institute of Information and Communication Technologies, Bulgarian Academy of Sciences (IICT-BAS)

For copies of reports, updates on project activities and other EuroMatrixPlus-related information, contact:

The EuroMatrixPlus Project Co-ordinator
Prof. Dr. Hans Uszkoreit, DFKI GmbH
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
uszkoreit@dfki.de
Phone +49 (681) 85775-5282 - Fax +49 (681) 85775-5338

Copies of reports and other material can also be accessed via the project's homepage: http://www.euromatrixplus.net/

Contents

Executive Summary **4**

1 WP1: Rich Tree-Based Statistical Translation **5**
1.1 Task 1.1: Shallow syntax modelling 5
1.2 Task 1.2: Develop rich contextual features 5
1.3 Task 1.3: TectoMT platform development 6
1.4 Task 1.4: Czech/English parallel data: extended annotation 6
1.4.1 PCEDT 2.0 6
1.4.2 PEDT 2.0 7
1.5 Task 1.5: Tree-based feature combination mapping 7
1.5.1 Motivation 7
1.5.2 Task Description 8
1.5.3 Data Description 8
1.5.4 Experiments Configuration 10
1.5.5 Conclusion 11
1.6 Task 1.6: Internal System Evaluation 11

References **14**

Executive Summary

This deliverable describes the work of WP1 (Rich Tree-Based Statistical Translation) of the EuroMatrixPlus project. The following pages provide more details on the progress in the following tasks:

Task	Months	Status
Task 1.1: Shallow syntax modeling	1–24	DONE
Task 1.2: Develop rich contextual features	12–24	DONE
Task 1.3: TectoMT platform development	1–36	DONE
Task 1.4: Czech-English annotation	1–33	DONE
Task 1.5: Tree-based features	1–24	DONE
Task 1.6: System evaluation	1–36	DONE

Work in Years Two and Three

Work was carried out on the following tasks as outlined in the Description of Work:

Task 1.1 *Shallow syntax modeling* (FBK, month 1–24)
We continued to carry out experiments with various system configurations and various languages; in particular, on Arabic a new effective verb reordering model was designed, implemented and experimentally tested. For details, see Section 1.1 and the cited publications.

Task 1.2 *Develop rich contextual features* (UEDIN, month 12-24)
We developed a number of frameworks that allow for the integration of a large number of features: work based on Gibbs sampling and SampleRank. We also re-implemented state-of-the-art approaches, namely MIRA and pairwise ranked optimization (PRO). These approaches have shown modest gain when using simple sparse features. We also explored the use of large contextual features in a maximum entropy approach during training (not tuning), aiming at reordering with nice gains over a strong baseline.

Task 1.3 *TectoMT platform development* (CU, month 1–36)
The Treex (formerly TectoMT) platform was greatly improved in terms of robustness (multiple new languages used and very large data for English and Czech processed), speed (profiling and a speedup of about 30%) as well as release. The core Treex modules are now publicly available from CPAN.

Task 1.4 *Czech-English annotation* (CU, month 1–33)
The annotation of both Czech and English data at the tectogrammatical layer of representation was finished in time. The data were wrapped and made publicly available in two separate releases: Prague Czech-English Dependency Treebank 2.0 (covering both sides of the treebank) and Prague English Dependency Treebank 2.0 (covering just the English tectogrammatical annotation with merged with some additional linguistic resources).

Task 1.5 *Tree-based features* (CU, month 1–24)
The aim of this task was to tackle the problem of predicting attributes of nodes in target-side deep-syntactic trees based on the source nodes. The study is presented in this deliverable. Unfortunately, the data-driven method suggested here does not reach satisfactory accuracy, so it was not incorporated in out deep-syntactic MT system (TectoMT, see also Task 1.3).

Task 1.6 *Internal system evaluation* (all participants, month 1–36)
Several partners have taken part in various evaluation campaigns, thus evaluating their systems in an open competition. Several studies of techniques of manual MT evaluation were also published.

Chapter 1

WP1: Rich Tree-Based Statistical Translation

1.1 Task 1.1: Shallow syntax modelling

Syntactic disfluencies in Arabic-to-English phrase-based SMT output are often due to incorrect verb reordering in VerbSubjectObject sentences. As a solution, we proposed (Bisazza and Federico, 2010; Bisazza et al., 2011) a chunk-based reordering technique to automatically displace clause-initial verbs in the Arabic side of a word-aligned parallel corpus. This method is used to preprocess the training data, and to collect statistics about verb movements. From this analysis we build specific verb reordering lattices on the test sentences before decoding, and test different lattice-weighting schemes. Finally, we train a feature-rich discriminative model to predict likely verb reorderings for a given Arabic sentence. The model scores are used to prune the reordering lattice, leading to better word reordering at decoding time. The application of our reordering methods to the training and test data resulted in consistent improvements on the NIST-MT 2009 ArabicEnglish benchmark, both in terms of BLEU (+1.06%) and of reordering quality (+0.85%) measured with the Kendall Reordering Score.

1.2 Task 1.2: Develop rich contextual features

The long-term goal of this task is to develop models for machine translation that may use arbitrary features over the source context of a word, phrase, sentence, and document. This involves both the development of models and training methods to allow for such rich featured models (machine learning research) and the investigation on which features are most beneficial (feature engineering research).

On the machine learning side, we have explored the use of Bayesian models that are trained on a sampling of the space of possible translations using Gibb's Sampling (Arun et al., 2009; Arun et al., 2010a; Arun et al., 2010b). This sampling is guaranteed to converge to the true distribution, and hence avoids the bias of just looking at the most likely events.

A different sampling method, SampleRank (Haddow et al., 2011), performs a random walk. While it does not come with the same guarantees, it tends to converge faster.

We also re-implemented MIRA (Hasler et al., 2011), which has been reported in the literature to work well with large tuning sets. We have shown that this implementation copes well with a large number of sparse features. We also re-implemented Pairwise Ranked Optimization (PRO), which gave us improvements when applied to a complex factored model.

Finally, we also explored the use of rich contextual features to aid reordering at the training stage. A maximum entropy classifier aids reordering decisions in a hierarchical model (Gao et al., 2011), improving over a strong baseline.

1.3 Task 1.3: TectoMT platform development

Treex (formerly TectoMT), which is a common platform developed for linguistically rich processing of text, went through a number of substantial design improvements in the last year. We focused especially on three aspects: (1) Robustness, (2) Speed, and (3) Support for external users:

Robustness: numerous tests were created for testing functionality correctness as well as for checking overall design and coding quality. Data-intensive tests were executed too: about 15 million sentence pairs from an English-Czech parallel corpus were analyzed by Treex tools, the same amount of English sentences were translated to Czech by Treex MT scenario, and Treex was also tested on a number of other languages (more than 30 treebanks are converted into Treex now).

Speed: a careful profiling of all core components was performed, which led to overall MT pipeline speed-up of about 30%.

Support for external users: all Treex core components are now fully documented and can be easily installed by anyone from CPAN, which is a broadly respected (de facto standard) repository of Perl libraries.

Besides implementing infrastructure improvements, Treex was used:

- in several NLP studies, such as Popel et al. (2011) Žabokrtský (2011) Mareček et al. (2011b)
- for building language data resources such as CzEng 1.0 (Bojar et al., 2011b), PCEDT 2.0 (Hajič et al., 2012), HamleDT (Mareček et al., 2011a).

1.4 Task 1.4: Czech/English parallel data: extended annotation

The annotation of Czech-English parallel data at the tectogrammatical (deep-syntactic) layer of annotation is completed. It proceeded internally organized into two independent but collaborating projects: PEDT (Prague English Dependency Treebank) for English and PCEDT_cz (Prague Czech-English Dependency Treebank, Czech side) for Czech. The corpus is being released by the Linguistic Data Consortium in two separate releases: the parallel Prague Czech-English Dependency Treebank 2.0 (also the Deliverable 1.2 of EuroMatrixPlus) and the Prague English Dependency Treebank 2.0.

1.4.1 PCEDT 2.0

The Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0[1]) is a major update of the Prague Czech-English Dependency Treebank 1.0 (Cuřín et al., 2004). It is a manually parsed Czech-English parallel corpus sized over 1.2 million running words in almost 50,000 sentences for each part.

The English part contains the entire Penn Treebank - Wall Street Journal Section (PTB, Marcus et al. (1999)). The Czech part consists of Czech translations of all of the Penn Treebank-WSJ texts. The corpus is 1:1 sentence-aligned. An additional automatic alignment on the node level (different for each annotation layer) is part of this release, too. The original Penn Treebank-like file structure (25 sections, each containing up to one hundred files) has been preserved. Only those PTB documents which have both POS and structural annotation (total of 2312 documents) have been translated to Czech and made part of this release.

Each language part is enhanced with a comprehensive manual linguistic annotation in the PDT 2.0 style (Hajič et al., 2006). The main features of this annotation style are:

[1] `http://ufal.mff.cuni.cz/pcedt2.0/`

- dependency structure of the content words and coordinating and similar structures (function words are attached as their attribute values),
- semantic labeling of content words and types of coordinating structures,
- argument structure, including an argument structure ("valency") lexicon for both languages,
- ellipsis and anaphora resolution.

This annotation style is called tectogrammatical annotation and it constitutes the tectogrammatical layer in the corpus. The most essential features of this annotation style have been specified in Hajič et al. (2012) and in the documentation accompanying the release.

1.4.2 PEDT 2.0

The English part alone comes as a separate release called PEDT 2.0[2], which is enhanced with the integrated visualization of other major annotation efforts performed by other teams, such as

- PropBank (Palmer et al., 2004),
- VerbNet[3] (Kipper et al., 2000),
- NomBank (Adam Meyers, 2008),
- flat noun phrase structures (by courtesy of D. Vadas[4] and J.R. Curran)
- BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005).

1.5 Task 1.5: Tree-based feature combination mapping

This section of the deliverable is devoted to predicting the values of node attributes in the target-side tectogrammatical tree when translating from the source tree. The section is a bit longer than other sections because the results are not available in any published paper so far.

1.5.1 Motivation

The translation over tectogrammatical layer (t-layer) includes the transfer of tectogrammatical nodes (t-nodes) from source-side tectogrammatical tree (t-tree) to the target-side one. This is a difficult task as it covers both the translation of the tree structure and the translation of the t-nodes' attributes. This chapter studies the latter problem.

The scope of this study is not limited to the translation over t-layer. The problem of predicting several possibly correlated target attributes from several also possibly correlated source attributes can be found in other methods of machine translation (e.g. Factored statistical MT (Koehn and Hoang, 2007)) and possibly also other domains.

The term of feature combination mapping sounds little fuzzy although the meaning we give it below is clear: Let us have a vector of nominal values $(s_1, ..., s_n)$ which is used for prediction of a vector of nominal values $(t_1, ..., t_n)$. Both source and target vector elements can be mutually dependent (we would say correlated in case of numerical values). The task is to find a way to predict the target vector elements so that they correspond to the source vector elements and at the same time they are mutually consistent.

The next section describes our particular task of predicting t-nodes which is an instance of this general problem. Section 1.5.3 describes our input data and conditions of the experiments,

[2] http://ufal.mff.cuni.cz/pedt2.0/

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

[4] http://sydney.edu.au/engineering/it/~dvadas1/

Section 1.5.4 the experiment configurations and results. Section 1.5.5 concludes the results and discusses possible future enhancements.

1.5.2 Task Description

The t-tree is a layer of linguistic annotation above the shallow syntactic tree. It omits the nodes with auxilliary words and adds the nodes for hidden (not directly expressed) parts of a sentence. An example of a sentence with its (manual) tectogrammatical annotation is shown in Figure 1.1.

The t-nodes in reality contain more attributes than shown in the figure. The complete list and descriptions of each attribut can be found in Mikulová et al. (2007).

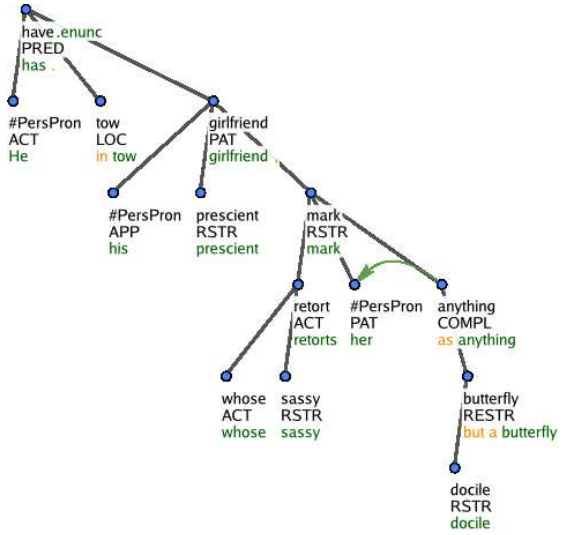

Figure 1.1: An example of tectogrammatical annotation. Original sentence: He has in tow his prescient girlfriend, whose sassy retorts mark her as anything but a docile butterfly.

Our task is to predict the English t-tree from the Czech t-tree and vice versa. As we do not go into tree structure prediction in this text, let us assume that we have a perfect 1-1 pairing of t-nodes on the source and target sides[5]. The remaining problem is to predict each node's target-side attributes. Our predictions are restricted to non-lexical attributes, i.e. do not consider the t-lemma on the source or target side.

1.5.3 Data Description

We use the automatic trees as available in the CzEng 0.9 corpus (Bojar and Žabokrtský, 2009) in so-called export format. The corpus is split by authors into ten parts. Eight of them are training data, the ninth part is a held-out set and the tenth part is the standard test data for CzEng 0.9. We use only the first part of the training data and separate one tenth of it by ourselves for testing.

We only use a subset of the selected data: we omit all the nodes which are not simply paired, i.e. the input data is composed of 1-1 pairs of t-nodes. We call these pairs "samples" in the following. The training and testing part consist of $4,282,033$ and $431,344$ samples, resp. The training part contains 426,700 unique Czech nodes (vectors of attributes) and $266,350$ unique English nodes (the t-lemma was already removed). This suggests that predicting the Czech nodes is a harder task, perhaps partly due to the fact that the automatic analysis of the Czech data is more fine-tuned. Statistics of individual attributes are given in Table 1.1.

[5]This is a viable assumption used in the TectoMT translation system (Popel et al., 2011), where the feature combination mapping is done using a mix of handcrafted and statistical rules. The assumption of 1-1 t-node mapping causes only 8% of translation errors of TectoMT (Popel and Žabokrtský, 2010)

Czech			English		
Attribute	Perplexity	Unique values	Attribute	Perplexity	Unique values
cs_aspect	1.82	4	en_aspect	1.00	1
cs_degcmp	1.55	4	en_degcmp	1.45	4
cs_deontmod	1.71	8	en_deontmod	1.78	7
cs_dispmod	1.73	3	en_dispmod	1.65	2
cs_formeme	18.04	1640	en_formeme	20.25	2493
cs_functor	12.95	67	en_functor	11.83	64
cs_gender	5.42	7	en_gender	1.51	5
cs_indeftype	1.22	12	en_indeftype	1.03	2
cs_is_clause_head	1.53	2	en_is_clause_head	1.50	2
cs_is_member	1.37	2	en_is_member	1.33	2
cs_is_passive	1.00	1	en_is_passive	1.09	2
cs_is_relclause_head	1.06	2	en_is_relclause_head	1.09	2
cs_iterativeness	1.60	3	en_iterativeness	1.65	2
cs_negation	1.91	3	en_negation	1.99	3
cs_nodetype	1.12	2	en_nodetype	1.25	5
cs_number	3.63	5	en_number	2.84	4
cs_numertype	1.23	6	en_numertype	1.00	1
cs_person	1.67	5	en_person	1.53	5
cs_politeness	1.19	3	en_politeness	1.00	1
cs_resultative	1.60	3	en_resultative	1.65	2
cs_sempos	7.27	20	en_sempos	5.48	11
cs_tense	1.94	5	en_tense	2.12	5
cs_val_frame_rf	1.00	1	en_val_frame_rf	1.00	1
cs_verbmod	1.80	5	en_verbmod	1.71	3

Table 1.1: Statistics of node attributes.

1.5.4 Experiments Configuration

We examine several ways to predict English t-node from Czech t-node and vice versa. We consider the usage of already predicted attributes in predicting. There are four experiment configurations which we run twice: once for Czech to English direction, once for English to Czech direction.

Le Zhang's maxent package[6] is trained on the training data and used for prediction. The absolute accuracy of the machine learning algorithm is not our primary goal. It could be definitely enhanced by proper tuning of the learning procedure. What we emphasize is the relative accuracy enhancement when demanding the consistency of the output target attributes. Maxent was chosen mainly for its good implementation which is able to work with this relatively large data collection. Its speed and memory efficiency outperformed the C5.0 software package[7] which we used in the early stage experimenting.

The `baseline` experiment is a prediction of the English-side attributes independently of each other. All the Czech attributes are used for each English attribute prediction.

The first modification (experiment `addrand`) over the baseline is keeping the already predicted values and adding them into the set of predictors. The order of predicted attributes is random but fixed throughout `addrand`. The training runs iteratively, the actual predicted values on training data of previous iterations are used when training the prediction of the next attribute.

Table 1.2 shows the comparison of accuracy when predicting individual attributes in `baseline` and `addrand` in direction from English to Czech and in direction from Czech to English, resp.

Czech			**English**		
Attribute	`baseline`	`addrand`	Attribute	`baseline`	`addrand`
cs_is_relclause_head	99.06 %	99.06 %	en_is_relclause_head	98.20 %	98.20 %
cs_person	95.02 %	95.01 %	en_person	96.78 %	96.77 %
cs_deontmod	94.61 %	94.61 %	en_deontmod	93.86 %	93.86 %
cs_nodetype	99.07 %	99.07 %	en_nodetype	97.03 %	97.02 %
cs_tense	91.44 %	91.44 %	en_tense	90.19 %	90.19 %
cs_degcmp	90.89 %	90.76 %	en_degcmp	90.90 %	90.89 %
cs_is_member	90.43 %	90.41 %	en_is_member	91.73 %	91.70 %
cs_functor	55.50 %	55.28 %	en_functor	55.41 %	55.15 %
cs_dispmod	93.62 %	93.56 %	en_dispmod	95.24 %	95.24 %
cs_is_passive	100.00 %	100.00 %	en_is_passive	98.25 %	98.25 %
cs_formeme	55.80 %	55.32 %	en_formeme	52.97 %	52.15 %
cs_aspect	89.85 %	89.68 %	en_aspect	100.00 %	100.00 %
cs_is_clause_head	94.39 %	94.42 %	en_is_clause_head	95.00 %	95.03 %
cs_resultative	95.91 %	95.84 %	en_resultative	95.24 %	95.25 %
cs_verbmod	93.15 %	93.05 %	en_verbmod	94.56 %	94.53 %
cs_gender	58.98 %	57.74 %	en_gender	96.75 %	96.72 %
cs_negation	93.41 %	92.87 %	en_negation	88.09 %	88.05 %
cs_politeness	96.70 %	96.64 %	en_politeness	100.00 %	100.00 %
cs_indeftype	97.27 %	97.27 %	en_indeftype	99.64 %	99.64 %
cs_val_frame_rf	100.00 %	100.00 %	en_val_frame_rf	100.00 %	100.00 %
cs_sempos	78.42 %	78.46 %	en_sempos	82.97 %	82.76 %
cs_number	76.56 %	76.33 %	en_number	81.98 %	79.75 %
cs_iterativeness	95.91 %	95.84 %	en_iterativeness	95.24 %	95.25 %
cs_numertype	98.44 %	98.45 %	en_numertype	100.00 %	100.00 %

Table 1.2: Accuracy of prediction of individual Czech and English attributes in experiments `baseline` and `addrand`.

[6] `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

[7] `http://www.rulequest.com/see5-info.html`

Attributes are listed in the same order as the prediction was made (random but fixed). In other words the prediction of `is_relclause_head` is the same task in both experiments, the prediction of `person` uses the predicted value of `is_relclause_head` in addition to source side attributes, the prediction of `deontmod` adds into the predictor set both `is_relclause_head` and `person` etc.

We use two metrics to express the overall accuracy of prediction:

1. the ratio of well-predicted attributes μ_{attr}
2. the ratio of well-predicted t-nodes μ_{node}

Tables 1.3 and 1.4 show the comparison of `baseline` and `addrand` according to these metrics.

Metric	`baseline`	`addrand`
μ_{nodes}	0.241	0.237
μ_{attrs}	0.894	0.893

Table 1.3: Overall metrics of Czech attributes accuracy.

Metric	`baseline`	`addrand`
μ_{nodes}	0.350	0.346
μ_{attrs}	0.917	0.915

Table 1.4: Overall metrics of English attributes accuracy.

Surprisingly, we can see that the baseline works better than our modification. Both metrics are slightly higher in the baseline experiment.

In the following experiments, we investigate if the accuracy of predicted nodes can be raised by choosing an appropriate order of predicting. Experiments `highlow` and `lowhigh` use two opposing orders of predictions: `highlow` predicts the attributes with the highest baseline accuracy first, whereas `lowhigh` predicts the attributes with the lowest baseline accuracy first. The comparison of their results is shown in Tables 1.5, 1.6 and 1.7.

We see that the overall accuracy of predicted nodes as wholes can be slightly raised by adding already predicted attributes to the set of predictors when predicting the following ones. The improvement is only achieved if the attributes are predicted in proper and somewhat surprising order: the attributes most difficult to predict have to be predicted first.

The general task of predicting vectors (of categorial values) from vectors is interesting on its own. Despite our efforts, we did not find any study of this specific machine learning task anywhere in the literature. The experiment on our particular dataset and the particular underlying machine learning algorithm suggests that the in sequential predicting, hard decisions should be made first. We believe that this conclusion may not hold if the ML technique performs some feature selection: the added values of attributes predicted in preceding steps should be automatically excluded if they tend to be misleading.

1.5.5 Conclusion

We have carried out experiments with predicting Czech and English attributes of t-layer nodes, aiming at the prediction of the whole vector of these attributes belonging to a node. The accuracy can be slightly increased if hard attributes are predicted first, and the guessed value is used for subsequent attributes.

We have not incorporated this fully automatic attribute prediction into the TectoMT system, because the overall performance is lower than the handcrafted system.

1.6 Task 1.6: Internal System Evaluation

CU has performed several studies of MT evaluation. Bojar (2011) analyzed two techniques of manual evaluation, the "sentence comprehension" (or also "monolingual post-editing") task as carried out in WMT10 and explicit marking of errors in MT outputs. Berka et al. (2011) evaluated four English-to-Czech MT systems using a another possible technique: posing comprehension questions to readers of MT output. Perhaps the most influential of our efforts in

Czech			English		
Attribute	lowhigh	highlow	Attribute	lowhigh	highlow
cs_functor	55.50 %	54.83 %	en_formeme	52.97 %	51.31 %
cs_formeme	55.69 %	54.74 %	en_functor	55.46 %	54.32 %
cs_gender	58.11 %	58.17 %	en_number	81.93 %	79.59 %
cs_number	76.52 %	76.26 %	en_sempos	83.03 %	82.77 %
cs_sempos	78.53 %	78.14 %	en_negation	88.09 %	87.96 %
cs_aspect	89.76 %	89.63 %	en_tense	90.20 %	90.11 %
cs_is_member	90.43 %	90.40 %	en_degcmp	90.76 %	90.64 %
cs_degcmp	90.85 %	90.61 %	en_is_member	91.72 %	91.65 %
cs_tense	91.38 %	91.39 %	en_deontmod	93.85 %	93.83 %
cs_verbmod	93.08 %	93.12 %	en_verbmod	94.53 %	94.51 %
cs_negation	93.10 %	92.62 %	en_is_clause_head	95.05 %	94.97 %
cs_dispmod	93.57 %	93.59 %	en_dispmod	95.26 %	95.24 %
cs_is_clause_head	94.36 %	94.39 %	en_iterativeness	95.26 %	95.24 %
cs_deontmod	94.53 %	94.61 %	en_resultative	95.26 %	95.24 %
cs_person	94.96 %	94.97 %	en_gender	96.68 %	96.76 %
cs_iterativeness	95.82 %	95.91 %	en_person	96.46 %	96.54 %
cs_resultative	95.83 %	95.91 %	en_nodetype	96.98 %	97.01 %
cs_politeness	96.59 %	96.68 %	en_is_relclause_head	98.213 %	98.21 %
cs_indeftype	97.34 %	97.27 %	en_is_passive	98.25 %	98.25 %
cs_numertype	98.45 %	98.41 %	en_indeftype	99.65 %	99.64 %
cs_is_relclause_head	99.06 %	99.06 %	en_aspect	100.00 %	100.00 %
cs_nodetype	99.05 %	99.07 %	en_numertype	100.00 %	100.00 %
cs_is_passive	100.00 %	100.00 %	en_politeness	100.00 %	100.00 %
cs_val_frame_rf	100.00 %	100.00 %	en_val_frame_rf	100.00 %	100.00 %

Table 1.5: Accuracy of prediction of individual Czech and English attributes in experiments `lowhigh` and `highlow`.

Metric	lowhigh	highlow
μ_{nodes}	**0.247**	0.234
μ_{attrs}	**0.895**	0.894

Table 1.6: Overall metrics of Czech attributes accuracy.

Metric	lowhigh	highlow
μ_{nodes}	**0.364**	0.342
μ_{attrs}	**0.917**	0.914

Table 1.7: Overall metrics of English attributes accuracy.

this respect was the collection and discussion of several technical issues with the current main WMT evaluation technique that relies on sentence ranking. This analysis is available in Bojar et al. (2011a).

In the last two years, FBK took part in the editions 2010 and 2011 of both IWSLT and WMT evaluation campaigns where the research pursued in EuroMatrixPlus could be publicly assessed.

In (Bisazza et al., 2010), new morphological segmentation rules were developed for Turkish-English. The combination of several Turkish segmentation schemes into a lattice input led to an improvement with respect to the previous year. The use of additional training data was explored for Arabic-English, while on the English to French task improvement was achieved over a strong baseline by automatically selecting relevant and high quality data from the available training corpora.

In WMT 2010, FBK participated to the machine translation shared task with phrase-based Statistical Machine Translation systems based on the Moses decoder for English-German and German-English translation (Hardmeier et al., 2010). The work focused on exploiting the available language modelling resources by using linear mixtures of large 6-gram language models and on addressing linguistic differences between English and German with methods based on word lattices. In particular, lattices were used to integrate a morphological analyzer for German into our system, and some initial work on rule-based word reordering was presented there.

FBK participated jointly with Uppsala University at the shared translation task of WMT 2011 (Hardmeier et al., 2011). Key features of the systems included anaphora resolution, hierarchical lexical reordering, data selection for language modelling, linear transduction grammars for word alignment and syntax-based decoding with monolingual dependency information.

In IWSLT 2011 evaluation campaign, FBK submitted runs in the English ASR track, the Arabic-English MT track and the English-French MT and SLT tracks (Ruiz et al., 2011). Concerning the MT and SLT systems, besides language specific pre-processing and the automatic introduction of punctuation in the ASR output, two major improvements are reported over the 2010 systems. First, we applied a fill-up method for phrase-table adaptation; second, we explored the use of hybrid class-based language models to better capture the language style of public speeches.

References

Catherine Macleod Adam Meyers, Ruth Reeves. 2008. NomBank v 1.0. LDC2008T23, ISBN: 1-58563-492-1.

Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez, and Philipp Koehn. 2009. Monte carlo inference and maximization for phrase-based translation. In *Conference on Computational Natural Language Learning*.

Abhishek Arun, Barry Haddow, and Philipp Koehn. 2010a. A unified approach to minimum risk training and decoding. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 365–374, Uppsala, Sweden, July. Association for Computational Linguistics.

Abhishek Arun, Barry Haddow, Philipp Koehn, Adam Lopez, Chris Dyer, and Miles Osborne. 2010b. Monte carlo techniques for phrase-based translation. *Machine Translation Journal*, 24(2):103–121.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77–86.

Arianna Bisazza and Marcello Federico. 2010. Chunk-Based Verb Reordering in VSO Sentences for Arabic-English Statistical Machine Translation. In *ACL Joint Workshop on SMT and MetricsMATR*, pages 241–249, Uppsala, Sweden.

Arianna Bisazza, Ioannis Klasinas, Mauro Cettolo, and Marcello Federico. 2010. FBK @ IWSLT 2010. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 53–58, Paris, France.

Arianna Bisazza, Daniele Pighin, and Marcello Federico. 2011. Chunk-Lattices for Verb Reordering in Arabic-English SMT. *Machine Translation: Special Issue on MT for Arabic*.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng 0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 93. under consideration.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011a. A grain of salt for the WMT manual evaluation. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, UK. Association for Computational Linguistics.

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2011b. Czeng 1.0.

Ondřej Bojar. 2011. Analyzing error types in english-czech machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:63–76.

Jan Cuřín, Martin Čmejrek, Jiří Havelka, Jan Hajič, Vladislav Kuboň, and Zdeněk Žabokrtský. 2004. Prague Czech-English Dependency Treebank Version 1.0. LDC2004T25, ISBN: 1-58563-321-6.

Yang Gao, Philipp Koehn, and Alexandra Birch. 2011. Soft dependency constraints for reordering in hierarchical phrase-based translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 857–868, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Barry Haddow, Abhishek Arun, and Philipp Koehn. 2011. Samplerank training for phrase-based machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 261–271, Edinburgh, Scotland, July. Association for Computational Linguistics.

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. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, Istanbul, Turkey, May. ELRA, European Language Resources Association. In print.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. 2006. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Christian Hardmeier, Arianna Bisazza, and Marcello Federico. 2010. FBK at WMT 2010: Word Lattices for Morphological Reduction and Chunk-Based Reordering. In *ACL Joint Workshop on SMT and MetricsMATR*, pages 88–92, Uppsala, Sweden.

Christian Hardmeier, Jörg Tiedemann, Markus Saers, Marcello Federico, and Mathur Prashant. 2011. The Uppsala-FBK systems at WMT 2011. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 372–378, Edinburgh, UK.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2011. Margin infused relaxed algorithm for Moses. *The Prague Bulletin of Mathematical Linguistics*, 96:69–78.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press / The MIT Press.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP-CoNLL*, pages 868–876. ACL.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Penn Treebank 3. LDC99T42, ISBN: 1-58563-163-9.

David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zeman, Zdeněk Žabokrtský, and Jan Hajič. 2011a. HamleDT - HArmonized multi-language dependency treebank.

David Mareček, Rudolf Rosa, Petra Galuščáková, and Ondřej Bojar. 2011b. Two-step translation with grammatical post-processing. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432, Edinburgh, UK. Association for Computational Linguistics.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Ševčíková, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, and Zdeněk Žabokrtský. 2007. Annotation on the tectogrammatical level in the prague dependency treebank. Technical Report 3.1, ÚFAL, Charles University.

Martha Palmer, Paul Kingsbury, Olga Babko-Malaya, Scott Cotton, and Benjamin Snyder. 2004. Proposition Bank I. LDC2004T14, ISBN: 1-58563-304-6.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Martin Popel, David Mareček, Nathan Green, and Zdeněk Žabokrtský. 2011. Influence of parser choice on dependency-based MT. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 433–439, Edinburgh, UK. Association for Computational Linguistics.

Nick Ruiz, Arianna Bisazza, Fabio Brugnara, Daniele Falavigna, Diego Giuliani, Suhel Jaber, Roberto Gretter, and Marcello Federico. 2011. FBK @ IWSLT 2011. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA.

Ralph Weischedel and Ada Brunstein. 2005. LDC2005T33, ISBN: 1-58563-362-3.

Zdeněk Žabokrtský. 2011. Treex an open-source framework for natural language processing.

In Markéta Lopatková, editor, *Information Technologies Applications and Theory*, volume 788, pages 7–14, Košice, Slovakia.