The Uppsala-FBK systems at WMT 2011

Christian Hardmeier
Jörg Tiedemann
Uppsala universitet
Inst. för lingvistik och filologi
Uppsala, Sweden
first.last@lingfil.uu.se

Markus Saers
Human Language
Technology Center
Hong Kong Univ. of
Science & Technology
masaers@cs.ust.hk

Marcello Federico
Mathur Prashant
Fondazione Bruno Kessler
Human Language Technologies
Trento, Italy
lastname@fbk.eu

Abstract

This paper presents our submissions to the shared translation task at WMT 2011. We created two largely independent systems for English-to-French and Haitian Creole-to-English translation to evaluate different features and components from our ongoing research on these language pairs. Key features of our systems include anaphora resolution, hierarchical lexical reordering, data selection for language modelling, linear transduction grammars for word alignment and syntax-based decoding with monolingual dependency information.

1 English to French

Our submission to the English-French task was a phrase-based Statistical Machine Translation based on the Moses decoder (Koehn et al., 2007). Phrase tables were separately trained on Europarl, news commentary and UN data and then linearly interpolated with uniform weights. For language modelling, we used 5-gram models trained with the IRSTLM toolkit (Federico et al., 2008) on the monolingual News corpus and parts of the English-French 10^9 corpus. More unusual features of our system included a special component to handle pronominal anaphora and the hierarchical lexical reordering model by Galley and Manning (2008). Selected features of our system will be discussed in depth in the following sections.

1.1 Handling pronominal anaphora

Pronominal anaphora is the use of pronominal expressions to refer to "something previously mentioned in the discourse" (Strube, 2006). It is a very common phenomenon found in almost all kinds of texts. Anaphora can be local to a sentence, or it can cross sentence boundaries. Standard SMT methods do not handle this phenomenon in a satisfactory way at present: For sentence-internal anaphora, they depend on the n-gram language model with its limited history, while cross-sentence anaphora is left to chance. We therefore added a word-dependency model (Hardmeier and Federico, 2010) to our system to handle anaphora explicitly.

Our processing of anaphoric pronouns follows the procedure outlined by Hardmeier and Federico (2010). We use the open-source coreference resolution system BART (Broscheit et al., 2010) to link pronouns to their antecedents in the text. Coreference links are handled differently depending on whether or not they cross sentence boundaries. If a coreference link points to a previous sentence, we process the sentence containing the antecedent with the SMT system and look up the translation of the antecedent in the translated output. If the coreference link is sentence-internal, the translation lookup is done dynamically by the decoder during search. In either case, the word-dependency model adds a feature function to the decoder score representing the probability of a particular pronoun choice given the translation of the antecedent.

In our English-French system, this model was only applied to the inanimate pronouns *it* and *they*, which seemed to be the most promising candidates for improvement since their French equivalents require gender marking. It was trained on data automatically annotated for anaphora taken from the news-commentary corpus, and the vocabulary of the predicted pronouns was limited to words recognised as pronouns by the POS tagger.

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 372–378,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

1.2 Hierarchical lexical reordering

The basic word order model of SMT penalises any divergence between the order of the words in the input sentence and the order of their translation equivalents in the MT output. All reordering must thus be driven by the language model when no other reordering model is present. Lexical reordering models making certain word order choices in the MT output conditional on the identity of the words involved have been a standard component in SMT for some years. The lexical reordering model usually employed in the Moses decoder was implemented by Koehn et al. (2005). Adopting the perspective of the SMT decoder, which produces the target sentence from left to right while covering source phrases in free order, the model distinguishes between three ordering classes, *monotone*, *swap* and *discontinuous*, depending on whether the source phrases giving rise to the two last target phrases emitted were adjacent in the same order, adjacent in swapped order or separated by other source words. Probabilities for each ordering class given source and target phrase are estimated from a word-aligned training corpus and integrated into MT decoding as extra feature functions.

In our submission, we used the hierarchical lexical reordering model proposed by Galley and Manning (2008) and recently implemented in the Moses decoder.[1] This model uses the same approach of classifying movements as *monotone*, *swap* or *discontinuous*, but unlike the phrase-based model, it does not require the source language phrases to be strictly adjacent in order to be counted as *monotone* or *swap*. Instead, a phrase can be recognised as adjacent to, or swapped with, a contiguous block of source words that has been segmented into multiple phrases. Contiguous phrase blocks are recognised by the decoder with a shift-reduce parsing algorithm. As a result, fewer jumps are labelled with the uninformative *discontinuous* class.

1.3 Data selection from the WMT Giga corpus

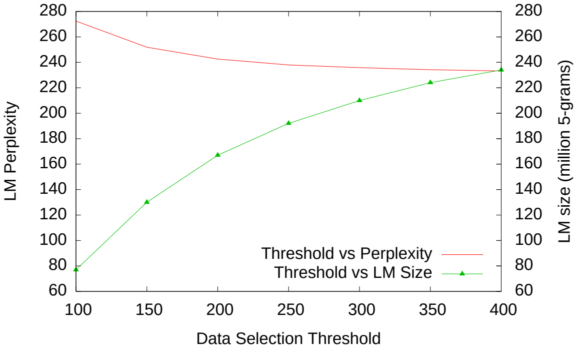


Figure 1: Perplexity and size of language models trained on data of the WMT Giga corpus that were selected using different perplexity thresholds.

One of the supplied language resources for this evaluation is the French-English WMT Giga corpus, aka 10^9 corpus, a large collection of parallel sentences crawled from Canadian and European Union sources. While this corpus was too large to be used for model training with the means at our disposal, we exploited it as a source of parallel data for translation model training as well as monolingual French data for the language model by filtering it down to a manageable size. In order to extract sentences close to the news translation task, we applied a simple data selection procedure based on perplexity. Sentence pairs were selected from the WMT Giga corpus if the perplexity of their French part with respect to a language model (LM) trained on French news data was below a given threshold. The rationale is that text sentences which are better predictable by the LM should be closer to the news domain. The threshold was set in a way to capture enough novel n-grams, from one side, but also to avoid adding too many irrelevant n-grams. It was tuned by training a 5-gram LM on the selected data and checking its size and its perplexity on a development set. In figure 1 we plot perplexity and size of the WMT Giga LM for different values of the data-selection threshold. Perplexities are computed on the newstest2009 set. As a good perplexity-size trade-off, the threshold 250 was chosen to estimate an additional 5-gram LM (WMT Giga 250) that was interpolated with the original News LM. The resulting improvement in perplexity is reported in table 1. For translation model data, a perplexity threshold of 159 was applied.

[1] The hierarchical lexical reordering model was implemented in Moses during MT Marathon 2010 by Christian Hardmeier, Gabriele Musillo, Nadi Tomeh, Ankit Srivastava, Sara Stymne and Marcello Federico.

LM	*Perplexity*	*OOV rate*
News	*146.84*	*0.82*
News + WMT Giga 250	*130.23*	*0.71*

Table 1: Perplexity reduction after interpolating the News LM with data selected from the 10^9 corpus.

	newstest		
	2009	2010	2011
Primary submission	0.246	0.286	0.284
w/o Anaphora handling	0.246	0.286	0.284
WMT Giga data			
w/o LM	0.244	0.289	0.280
w/o TM	0.247	0.286	0.282
w/o LM and TM	0.247	0.289	0.278
Lexical reordering			
phrase-based reo	0.239	0.281	0.275
no lexical reo	0.239	0.281	0.275
with LDC data	0.254	0.293	0.291

Table 2: Ablation test results (case-sensitive BLEU)

1.4 Results and Ablation tests

Owing to time constraints, we were not able to run thorough tests on our system before submitting it to the evaluation campaign. We therefore evaluated the various components included in a *post hoc* fashion by running ablation tests. In each test, we left out one of the system components to identify its effect on the overall performance. The results of these tests are reported in table 2.

Performance-wise, the most important particularity of our SMT system was the hierarchical lexical reordering model, which led to a sizeable improvement of 0.7, 0.5 and 0.9 BLEU points for the 2009, 2010 and 2011 test sets, respectively. We had previously seen negative results when trying to apply the same model to English-German SMT, so its performance seems to be strongly dependent on the language pair it is used with.

Compared to the scores obtained using the full system, the anaphora handling system did not have any effect on the BLEU scores. This result is similar to our result for English-German translation (Hardmeier and Federico, 2010). Unfortunately, for English-French, the negative results extends to the pronoun translation scores (not reported here), where slightly higher recall with the word-dependency model was overcompensated by degraded precision, so the outcome of the experiments clearly suggests that the anaphora handling procedure is in need of improvement.

The effect of the WMT Giga language model differs among the test sets. For the 2009 and 2011 test sets, it results in an improvement of 0.2 and 0.4 BLEU points, respectively, while the 2010 test set fares better without this additional language model. However, it should be noted that there may be a problem with the 2010 test set and the News language model, which was used as a component in all our systems. In particular, upgrading the News LM data from last year's to this year's release led to an improvement of 4 BLEU points on the 2010 test set and an unrealistically low perplexity of 73 as compared to 130 for the 2009 test set, which makes us suspect that the latest News LM data may be tainted with data from the 2010 test corpus. If this is the case, the 2010 test set should be considered unreliable for LM evaluation. The benefit of adding WMT Giga data to the translation model is less clear. For the 2009 and 2010 test sets, this leads to a slight degradation, but for the 2011 corpus, we obtained a small improvement.

Our shared task submission did not use the French Gigaword corpus from the Linguistic Data Consortium (LDC2009T28), which is not freely available to sites without LDC membership. After the submission, we ran a contrastive experiment including a 5-gram model trained on this corpus, which led to a sizeable improvement of 0.7–0.8 BLEU points across all test sets.

2 Haitian Creole to English

Our experiments with the Haitian Creole-English data are independent of the system presented for the English to French task above. We experimented with both phrase-based SMT and syntax-based SMT. The main questions we investigated were i) whether we can improve word alignment and phrase extraction for phrase-based SMT and ii) whether we can integrate dependency parsing into a syntax-based approach. All our experiments were conducted on the *clean* data set using Moses for training and decoding. In the following we will first describe the experiments with phrase-based models and linear transduction grammars for word alignment and, thereafter, our findings from integrating English dependency parses into a syntax-based approach.

2.1 Phrase-based SMT

The phrase-based system that we used in this series of experiments uses a rather traditional setup. For the translations into English we used the news data provided for the other translations tasks in WMT 2011 to build a large scale-background language model. The English data from the Haitian Creole task were used as a separate domain-specific language model. For the other translation direction we only used the in-domain data provided. We used standard 5-gram models with Witten-Bell discounting and backoff interpolation for all language models. For the translation model we applied standard techniques and settings for phrase extraction and score estimations. However, we applied two different systems for word alignment: One is the standard GIZA++ toolbox implementing the IBM alignment models (Och and Ney, 2003) and extensions and the other is based on transduction grammars which will briefly be introduced in the next section.

2.1.1 Alignment with PLITGs

By making the assumption that the parallel corpus constitutes a *linear transduction* (Saers, 2011)[2] we can induce a grammar that is the most likely to have generated the observed corpus. The grammar induced will generate a parse forest for each sentence pair in the corpus, and each parse tree in that forest will correspond to an alignment between the two sentences. Following Saers et al. (2010), the alignment corresponding to the best parse can be extracted and used instead of other word alignment approaches such as GIZA++. There are several grammar types that generate linear transductions, and in this work, *stochastic bracketing preterminalized linear inversion transduction grammars* (PLITG) were used (Saers and Wu, 2011). Since we were mainly interested in the word alignments, we did not induce phrasal grammars.

Although alignments from PLITGs may not reach the same level of translation quality as GIZA++, they make different mistakes, so both complement each other. By duplicating the training corpus and aligning each copy of the corpus with a different alignment tool, the phrase extractor seems to be able to pick the best of both worlds, producing a phrase table that is superior to one produced with either of the alignments tools used in isolation.

2.1.2 Results

In the following we present our results on the provided test set[3] for translating into both languages with phrase-based systems trained on different word alignments. Table 3 summarises the BLEU scores obtained.

English-Haitian	BLEU	phrase-table
GIZA++	0.2567	3,060,486
PLITG	0.2407	5,007,254
GIZA++ & PLITG	**0.2572**	7,521,754
Haitian-English	BLEU	phrase-table
GIZA++	0.3045	3,060,486
PLITG	0.2922	5,049,280
GIZA++ & PLITG	**0.3105**	7,561,043

Table 3: Phrase-based SMT (*pbsmt*) on the Haitian Creole-English test set with different word alignments.

From the table we can see that phrase-based systems trained on PLITG alignments performs slightly worse than the ones trained on GIZA++. However combining both alignments with the simple data duplication technique mentioned earlier produces the overall best scores in both translation directions. The fact that both alignments lead to complementary information can be seen in the size of the phrase tables extracted (see table 3).

2.2 Syntax-based SMT

We used Moses and its syntax-mode for our experiments with hierarchical phrase-based and syntax-augmented models. Our main interest was to investigate the influence of monolingual parsing on the translation performance. In particular, we tried to integrate English dependency parses created by MaltParser (Nivre et al., 2007) trained on the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993) extended with about 4000 questions

[2] A transduction is a set of pairs of strings, and thus represents a relation between two languages.

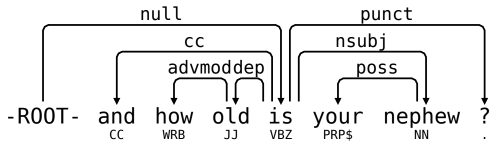
[3] We actually swapped the development set and the test set by mistake. But, of course, we never mixed development and test data in any result reported.

from the Question Bank (Judge et al., 2006). The conversion to dependency trees was done using the Stanford Parser (de Marneffe et al., 2006). Again, we ran both translation directions to test our settings in more than just one task. Interesting here is also the question whether there are significant differences when integrating monolingual parses on the source or on the target side.

The motivation for applying dependency parsing in our experiments is to use the specific information carried by dependency relations. Dependency structures encode functional relations between words that can be seen as an interface to the semantics of a sentence. This information is usually not available in phrase-structure representations. We believe that this type of information can be beneficial for machine translation. For example, knowing that a noun acts as the subject of a sentence is more informative than just marking it as part of a noun phrase. Whether or not this information can be explored by current syntax-based machine translation approaches that are optimised for phrase-structure representations is a question that we liked to investigate. For comparison we also trained hierarchical phrase-based models without any additional annotation.

2.2.1 Converting projective dependency trees

First we needed to convert dependency parses to a tree representation in order to use our data in the standard models of syntax-based models implemented in Moses. In our experiments, we used a parser model that creates projective dependency graphs that can be converted into tree structures of nested segments. We used the yield of each word (referring to that word and its transitive dependents) to define spans of phrases and their dependency relations are used as span labels. Furthermore, we also defined pre-terminal nodes that encode the part-of-speech information of each word. These tags were obtained using the HunPos tagger (Halácsy et al., 2007) trained on the Wall Street Journal section of the Penn Treebank. Figure 2 illustrates the conversion process. Tagging and parsing is done for all English data without any manual corrections or optimisation of parameters. After the conversion, we were able to use the standard training procedures implemented in Moses.



```
<tree label="null">
  <tree label="cc">
    <tree label="CC">and</tree>
  </tree>
  <tree label="dep">
    <tree label="advmod">
      <tree label="WRB">how</tree>
    </tree>
    <tree label="JJ">old</tree>
  </tree>
  <tree label="VBZ">is</tree>
  <tree label="nsubj">
    <tree label="poss">
      <tree label="PRP$">your</tree>
    </tree>
    <tree label="NN">nephew</tree>
  </tree>
  <tree label="punct">
    <tree label=".">?</tree>
  </tree>
</tree>
```

Figure 2: A dependency graph from the training corpus and its conversion to a nested tree structure. The yield of each word in the sentence defines a span with the label taken from the relation of that word to its head. Part-of-speech tags are used as additional pre-terminal nodes.

2.2.2 Experimental Results

We ran several experiments with slightly different settings. We used the same basic setup for all of them including the same language models and GIZA++ word alignments that we have used for the phrase-based models already. Further, we used Moses for extracting rules of the syntax-based translation model. We use standard settings for the baseline system (=hiero) that does not employ any linguistic markup. For the models that include dependency-based trees we changed the maximum span threshold to a high value of 999 (default: 15) in order to extract as many rules as possible. This large degree of freedom is possible due to the otherwise strong constraints on rule flexibility imposed by the monolingual syntactic markup. Rule tables are dramatically smaller than for the unrestricted hierarchical models (see table 4).

However, rule restriction by linguistic constraints usually hurts performance due to the decreased coverage of the rule set. One common way of improving

reference	Are you going to let us die on Ile à Vaches which is located close the city of Les Cayes. I am ...
pbsmt	Do you are letting us die in Ilavach island's on in Les Cayes. I am ...
hiero	do you will let us die in the island Ilavach on the in Les Cayes . I am ...
samt2	Are you going to let us die in the island Ilavach the which is on the Les. My name is ...
reference	I'm begging you please help me my situation is very critical.
pbsmt	Please help me please. Because my critical situation very much.
hiero	please , please help me because my critical situation very much .
samt2	Please help me because my situation very critical.
reference	I don't have money to go and give blood in Port au Prince from La Gonave.
pbsmt	I don't have money, so that I go to give blood Port-au-Prince since lagonave.
hiero	I don 't have any money , for me to go to give blood Port-au-Prince since lagonave .
samt2	I don't have any money, to be able to go to give blood Port-au-Prince since Gonâve Island.

Figure 3: Example translations for various models.

English-Haitian	BLEU	number of rules
hiero	**0.2549**	34,118,622
malt (source)	0.2180	1,628,496
- binarised	0.2327	9,063,933
- samt1	0.2311	11,691,279
- samt2	0.2366	29,783,694
Haitian-English	BLEU	number of rules
hiero	**0.3034**	33,231,535
malt (target)	0.2739	1,922,688
- binarised	0.2857	8,922,343
- samt1	0.2952	11,073,764
- samt2	*0.2954*	*24,554,317*

Table 4: Syntax-based SMT on the Haitian Creole-English test set with (=malt) or without (=hiero) English parse trees and various parse relaxation strategies. The final system submitted to WMT11 is *malt(target)-samt2*.

rule extraction is based on tree manipulation and relaxed extraction algorithms. Moses implements several algorithms that have been proposed in the literature. Tree binarisation is one of them. This can be done in a left-branching and in a right-branching mode. We used a combination of both in the settings denoted as *binarised*. The other relaxation algorithms are based on methods proposed for syntax-augmented machine translation (Zollmann et al., 2008). We used two of them: *samt1* combines pairs of neighbouring children nodes into combined complex nodes and creates additional complex nodes of all children nodes except the first child and similar complex nodes for all but the last child. *samt2* combines any pair of neighbouring nodes even if they are not children of the same parent. All of these relaxation algorithms lead to increased rule sets (table 4). In terms of translation performance there seems to be a strong correlation between rule table size and translation quality as measured by BLEU. None of the dependency-based models beats the unrestricted hierarchical model. Both translation directions behave similar with slightly worse performances of the dependency-based models (relative to the baseline) when syntax is used on the source language side. Note also that all syntax-based models (including hiero) are below the corresponding phrase-based SMT systems. Of course, automatic evaluation has its limits and interesting qualitative differences may be more visible in manual assessments. The use of linguistic information certainly has an impact on the translation hypotheses produced as we can see in the examples in figure 3. In the future, we plan to investigate the effect of dependency information on grammaticality of translated sentences in more detail.

3 Conclusions

In our English-French and Haitian Creole-English shared task submissions, we investigated the use of anaphora resolution, hierarchical lexical reordering and data selection for language modelling (English-French) as well as LTG word alignment and syntax-based decoding with dependency information (Haitian Creole-English). While the results for the systems with anaphora handling were somewhat disappointing and the effect of data filtering was inconsistent, hierarchical lexical reordering brought substantial improvements. We also obtained consistent gains by combining information from different word aligners, and we presented a simple way of including dependency parses in standard tree-based decoding.

Acknowledgements

Most of the features used in our English-French system were originally developed while Christian Hardmeier was at FBK. Activities at FBK were supported by the EuroMatrixPlus project (IST-231720) and the T4ME network of excellence (IST-249119), both funded by the DG INFSO of the European Commission through the Seventh Framework Programme.

References

Samuel Broscheit, Massimo Poesio, Simone Paolo Ponzetto, Kepa Joseba Rodriguez, Lorenza Romano, Olga Uryupina, Yannick Versley, and Roberto Zanoli. 2010. BART: A multilingual anaphora resolution system. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech 2008*, pages 1618–1621. ISCA.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 847–855, Honolulu, Hawaii, October. Association for Computational Linguistics.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Hunpos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 209–212.

Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 283–289.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 497–504.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, et al. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *International workshop on spoken language translation*, Pittsburgh.

Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: open source toolkit for Statistical Machine Translation. In *Annual meeting of the Association for Computational Linguistics: Demonstration session*, pages 177–180, Prague.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19:313–330, June.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29:19–51.

Markus Saers and Dekai Wu. 2011. Principled induction of phrasal bilexica. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium, May.

Markus Saers, Joakim Nivre, and Dekai Wu. 2010. Word alignment with stochastic bracketing linear inversion transduction grammar. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 341–344, Los Angeles, California, June.

Markus Saers. 2011. *Translation as Linear Transduction: Models and Algorithms for Efficient Learning in Statistical Machine Translation*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology.

M. Strube. 2006. Anaphora and coreference resolution, Statistical. In *Encyclopedia of language and linguistics*, pages 216–222. Elsevier.

Andreas Zollmann, Ashish Venugopal, Franz Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical mt. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pages 1145–1152.