D 6.1: Improved confidence estimation and hybrid architectures for machine translation

Yu Chen, Andreas Eisele, Christian Federmann, Michael Jellinghaus, Silke Theison

Distribution: Public

EuroMatrix
Statistical and Hybrid Machine Translation
Between All European Languages
IST 034291 Deliverable D 6.1

December 2007, version 2

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

Project ref no.	IST-034291
Project acronym	EuroMatrix
Project full title	Statistical and Hybrid Machine Translation Between All European Languages
Instrument	STREP
Thematic Priority	Information Society Technologies
Start date / duration	01 September 2006 / 30 Months

Distribution	Public
Contractual date of delivery	December 2007
Actual date of delivery	December 2007
Deliverable number	D 6.1
Deliverable title	Improved confidence estimation and hybrid architectures for machine translation
Type	
Status & version	
Number of pages	28
Contributing WP(s)	6
WP / Task responsible	Andreas Eisele
Other contributors	
Author(s)	Yu Chen, Andreas Eisele, Christian Federmann, Michael Jellinghaus, Silke Theison
EC project officer	Xavier Gros
Keywords	

The partners in EuroMatrix are: Saarland University (USAAR)
University of Edinburgh (UEDIN)
Charles University (CUNI-MFF)
CELCT
GROUP Technologies
MorphoLogic

For copies of reports, updates on project activities and other EuroMatrix-related information, contact:
The EuroMatrix Project Co-ordinator
Prof. Hans Uszkoreit
Universität des Saarlandes, Computerlinguistik
Postfach 15 11 50
66041 Saarbrücken, Germany
uszkoreit@coli.uni-sb.de
Phone +49 (681) 302-4115- Fax +49 (681) 302-4700

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

Contents

Chapter 1

Introduction

This document describes work done in work package WP6 of the EuroMatrix project towards the definition and implementation of improved confidence estimates for the quality of machine translation (MT) output. It explains a certain deviation from the original workplan and presents the results achieved so far in all the subtasks of the work package.

1.1 Motivation

Improved confidence estimates are an essential ingredience of hybrid and multi-engine MT systems, as they are needed when choosing between alternative translations produced by different MT engines. Moreover, the usual statistical MT (SMT) architecture relies on models for both translation adequacy and fluency, so any improvement in these models will have a beneficial effect on the outcome of SMT.

Improved models of translation quality can also play an important role in the deployment of MT technology in professional applications, where the effort (and cost) of post-editing MT output has to be compared to the effort of providing a human translation from scratch. A system that is aware of its own weaknesses can alleviate the burden of a professional translator who is confronted with MT output; only if the output is promising, the translator should spend time looking at it.

1.2 Deviation from Workplan

The workplan for WP6 defines the following tasks and deliverables:

Original outline of WP6

Task 6.1: Aligning translation candidates with input
Integrating translation candidates from multiple engines requires knowledge about the relation of candidate substrings to substrings of the source text. For MT engines that do not deliver this kind of information, this alignment needs to be computed externally. Variants of standard algorithms for word- and phrase alignment will be implemented so solve this subtask in a generic and reliable way.
Task 6.2: Develop confidence estimates

Various kinds of linguistic knowledge and statistical evidence need to be integrated in order to improve automatically computed estimated of translation quality. For this integration, state-of-the-art machine learning techniques as well as novel types of training data will be employed.

Task 6.3: Implement improved quality estimates

Build a modular implementation of improved quality estimators that make them accessible both for confidence estimation within a hybrid MT engine and as a standalone module, to be used e.g. for MT evaluation.

Task 6.4: Develop hybrid system and participate in second evaluation campaign

A hybrid system architecture will be developed that combines quality estimation, word alignment and a varying number of MT and other linguistic modules, both black or white boxes, into a unified system that maximizes the benefit from complementary advantages of its ingredient modules.

Deliverables

Number	Title	Month due	Type	Dissemination
D6.1	Implementation of improved confidence metrics	15	RS,C	public
D6.2	Integrated hybrid system	27	S,C	public

D6.1 Implementation of improved confidence metrics

Confidence metrics for MT output will be implemented as documented APIs that allow to use these metrics as a module in a hybrid MT setup.

D6.2 Integrated hybrid system

A hybrid MT system incorporating rule-based and corpus-based functionality from the engines in WPs 2..4 will be built and made available via the EuroMatrix website.

Dependencies among tasks

The dependencies between the tasks defined in the workplan are such that Task 6.4 depends on Tasks 6.1 and 6.3, whereas Task 6.3 depends on Task 6.2. A logical order of the steps to be done would therefore complete Task 6.3 before Task 6.4 would be started, and for that reason deliverable D6.1, due in month 15, was defined to focus on the completion of Tasks 6.2 and 6.3. However, it turned out during the course of the work that this strict chronological order would not allow feed-back from the evaluation of the integrated system to influence the implementation of Tasks 6.2 and 6.3, and that it would be better to come up with a completed hybrid setup earlier in the project, with baseline implementations of the quality metrics taking the place of the completed implementation of Task 6.3. After changing the order of the tasks accordingly are now working backwards, from a completed end-to-end setup involving a baseline module for quality estimation towards an improved implementation of the quality estimates that is being optimized in the light of the outcome of evaluation results.

Chapter 2

Improved models of translation quality

Whereas evaluation of translation quality is an area with a long history and a huge number of publications (see also the list of references given in Deliverable 3.1), we are here concerned with the special case of fully automatic evaluation of the quality of translation candidates generated in a hybrid MT setup involving multiple MT engines. The most important difference between confidence estimation and automatic evaluation is the fact that confidence estimation lacks the reference human translations that are normally employed in automatic evaluation.

It is easy to see that in the general case, a highly reliable method for estimating translation quality cannot be much easier than the MT problem itself; if we had access to automatic ways to perfectly predict translation quality, we could generate, at least in principle, perfect translations – the task of fully automatic high-quality MT would boil down to the task of making the naive generate and test approach more efficient. Hence, we have to restrict ourselves to more modest goals, which can be characterized as distinguishing really good and really bad candidates with some reliability. Good translations need to satisfy two criteria; one one hand, they need to be close to the original text and convey the same meaning or should at least be able to fulfill the same function, while on the other hand they should be rendered as well-formed utterances of the target language. The problem can hence be naturally decomposed into two parts, namely the estimation of adequacy and of fluency[1].

2.1 Models of fluency

One important part of the overall task of estimating translation quality is the judgement whether some piece of text generated by an MT system is a well-formed and plausible utterance of the target language.

[1]It is not by chance that these are also the main criteria used in recent human evaluation campaigns organized by NIST and by other groups. However, as it has turned out that in practice, the performance of systems on both scales is highly correlated, more recent evaluation campaigns try to ease the task fo the evaluators by collapsing both criteria. In the case of automatic evaluation, there are no strong reasons for similar moves.

For this judgement, two different knowledge sources can be employed. On one hand, one can test the grammaticality of the proposed string using linguistic knowledge about the target language. On the other hand, one can test whether the proposed utterance follows typical patterns of language use in the target language.

We informally investigated 2000 sentences translated by the EN → DE branch of the SMT system described in (Koehn, 2005)[2]. We compared these translations with translations obtained from the rule-based MT system *translate pro*, Version 8, sold by Digital Publishing.

This comparison made it quite clear that for this language pair both types of systems generate rather different patterns of typical errors. Whereas SMT was generally rather good in lexical choice, the number of syntactic errors was significantly higher than for the rule-based system. The rule-based system was able to come up with syntactically well-formed translations more often than the SMT system, but had a significantly higher rate of wrong disambiguations.

Fig. 2.1 gives some typical examples of translations proposed by the two systems, as well as a short description of the problems they contain. A more exhaustive investigation with a quantitative evaluation of different error types has not been done on this data set, but is being prepared for more recent versions of the MT systems.

2.1.1 Statistical language models

As motivated above, we included a statistical model of fluency into the hybrid system used in the experiments performed so far. As explained in Chapter 3 of this document, one of the hybrid architectures being developed relies on the SMT decoder Moses for the integration of outcomes of multiple MT engines. So far, this decoder uses standard nGram-based language models, and actually offers users the choice among several implementations that are essentially equivalent with respect to the probabilities that are assigned to translation candidates ((Stolcke, 2002) or (Federico and Cettolo, 2007)).

In order to gain experience with alternative language models, one of the authors of this document worked on an implementation of statistical language models that is suitable to be trained on very large repositories, such as the huge repository of 5-gram counts published by Google in 2006 [3]. The details of this work are given in a diploma thesis (Federmann, 2007), which also has links to the full source code developed as part of this work.

2.1.2 Grammar-based language models

As can be seen from the examples given in Fig. 2.1, statistical language models based on nGrams on the level of words do not seem capable to detect certain types of important linguistic constraints, including the constraint that a well-formed sentence typcially has one inflected verb at an appropriate position, and the constraint that the endings of inflected words should agree among themselves and with the syntactic function the larger phrase plays in the sentence.

[2]Outcomes for 2000 test sentences can be obtained from `http://www.statmt.org/matrix/` for all 110 language pairs by clicking on one of the cells of the matrix.

[3]Distributed by the Linguistic Data Consortium as LDC2006T13

Original	*I wish the negotiators continued success with their work in this important area.*	
SMT	*Ich wünsche de***r** *Verhandlungsführer_ fortgesetzte_ Erfolg bei ihrer Arbeit in diesem wichtigen Bereich.*	wrong inflectional endings
RBMT	*Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich* **fortsetzten**.	wrong disambiguation

Original	*We seem sometimes to have lost sight of this fact.*	
SMT	*Manchmal scheinen wir aus den Augen verloren haben,* **diese Tatsache.**	unnatural word order
RBMT	*Wir scheinen manchmal* **Anblick** *dieser Tatsache verloren zu haben.*	idiomatic expression not known

Original	*I would like to close with a procedural motion.*	
SMT	*Ich möchte abschließend eine Frage zur Geschäftsordnung _.*	verb missing
RBMT	*Ich möchte mit einer* **verfahrenstechnischen Bewegung** *schließen.*	technical term not known

Original	*The leaders of Europe have not formulated a clear vision.*	
SMT	*Die Führung Europas* **nicht formuliert** *eine klare Vision.*	unnatural word order
RBMT	*Die* **Leiter von Europa** *haben keine klare Vision formuliert.*	unnatural expression

Figure 2.1: Typical errors in statistical and rule-based MT

Ideally, one would like to make use of linguistic constraints and a full-fledged syntactic analysis to detect errors of this type. However, currently even the best MT output will typically contain a significant number of errors, so it would not be wise to expect complete syntactic correctness. Instead, one would like to prefer candidates that minimize the number of syntactic errors or preferrably take the number of errors, weighted by severity, into account when selecting good translations.

We have started to work on this topic along two lines.

Recent work on increasing the robustness of deep linguistic analysis based on head-driven phrase structure grammars (HPSG) has lead to significant progress in that area, as documented in (Zhang, 2007), which makes is now realistic to employ a robust deep parser as one of the modules used in the estimation of grammaticality. We have started experiments to use this robust parser as one of the knowledge sources to obtain quality judgements, but as the experiments are still ongoing, it is still too early to report details. Whereas the work reported in (Zhang, 2007) is focused on English, we see the biggest potential for grammar-based language modeling for more richly inflected languages. Work on porting these techniques to German are on the way and will be used for experiments in the second half of the project.

Even without full-fledged syntactic analysis, many problems with the grammaticality of utterances can be detected by applying suitable linguistic rules that may be simpler and inherently more robust. Using a simple part-of-speech tagger it is possible to identify one of the most frequent problems of a purely statistical MT system, the fact that the output of such as system often lacks an inflected verb. More sophisticated rule systems aimed at the identification of errors have been developed to support technical authoring via controlled language. Although such systems are not designed to detect errors in MT output (rather they are more often used for controlling the input given to MT systems), it seems conceivable that problems reported by such tools will be helpful to distinguish good and bad parts within MT output. The company Acrolinx GmbH, a spin-off companies of DFKI, has implemented Acrocheck^{TM}, a software solution for quality assurance in technical documentation, which is able to check style, terminology, consistency, grammar, and spelling in technical documents for English, German, and French.

The nexts steps towards grammar-based estimation of translation quality will involve the two approaches sketched above, together with simpler models such as part-of-speech taggers and spotting of unknown or rare words in the target language. We will use the data collected in the shared task of the 2nd ACL workshop on SMT (Callison-Burch et al., 2007), together with human ratings, to automatically learn a mapping from problems detected by these tools to approximate ratings of text quality.

2.2 Improved models of adequacy

2.2.1 Improving precision

Even if the output of a MT system was a perfect utterance of the target language, this would of course not yet qualify it as a useful translation of a given source text, unless it can be shown that the two texts share their meaning or can at least achieve the same function in the intended use of the texts. Tools like the on-line translation service

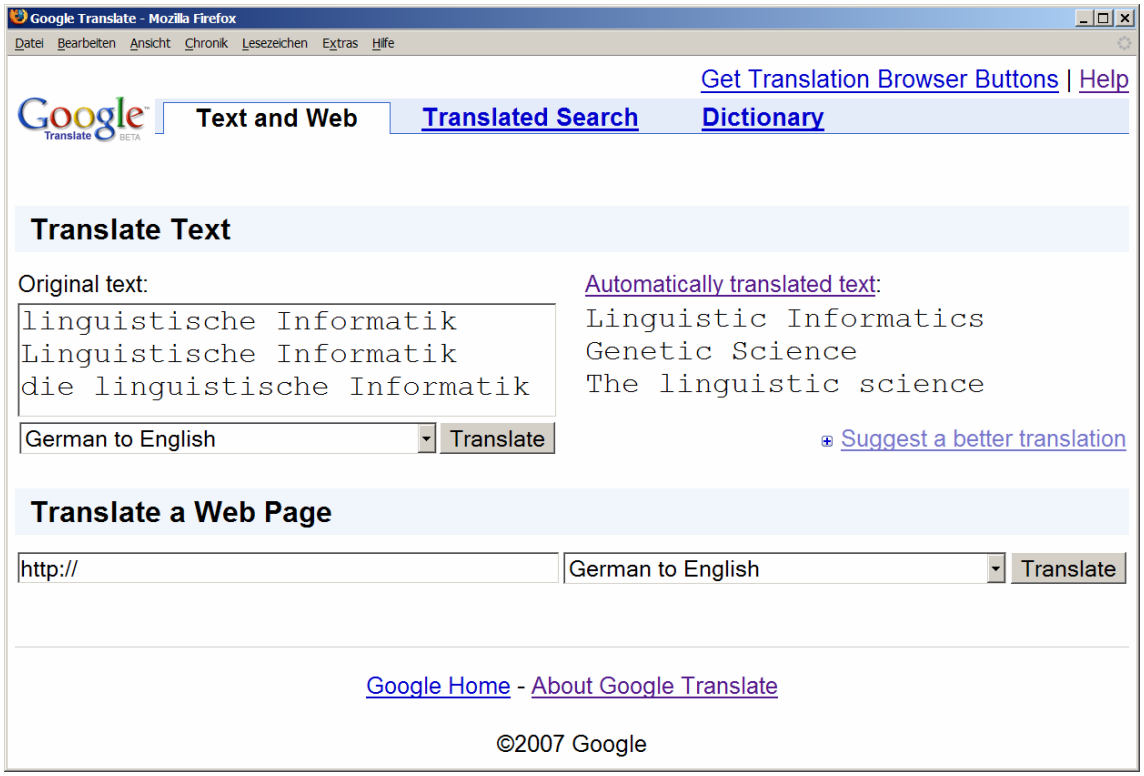

Figure 2.2: Some errors in Google's MT service

of Google, which has recently been adapted to use SMT technology for all language pairs, sometimes produce severe errors which can be quite hard to spot, as the generated utterances are well-formed and sound natural, but have little to do with the input.

Fig. 2.2 shows Google's tool generating three completely different translations for the same input phrase, two of which are completely wrong[4] and the choice between which depend on little details like capitalization of a word or the presence of a determiner.

Such examples lead to the question for the cause of these and similar errors and whether there are ways to automatically detect them.

When trying to quantify the quality of a proposed translation of an utterance or when comparing several competing proposals, it is important to note that statistical models of correspondences between phrases and substrings that are used within the SMT framework to generate translations can also be used to assess the quality of translations generated by other mechanisms. Hence, improving the translation models can have two positive effects at the same time: In a SMT setup, generating translations using better translation model will help a system to focus the computation on proposals that are better on average, and hence make better use of the available computational resources. In a hybrid setup, using SMT models to estimate the quality of RBMT output can be helpful to spot errors in proposals generated by the RBMT engine, but they can also point the developers to situations where the SMT models lack coverage compared to the RBMT engine.

[4]The snapshot was taken in November 2007; at the end of December 2007 the errors still persisted, but have been fixed in the meantime.

Errors like those shown in Fig. 2.2 can creep in by automatic extraction of corresponding pairs of expressions that are not mutual translations. This can be caused by the fact that documents or parts of the documents used as training data may not be mutual translations in the first place, or, if they are, by errors in the automatically computed alignments.

We do not know of any fully automatic way to avoid these types of errors, but we can see several approaches that may be helpful to detect them:

- Low frequencies are generally a sign of reduced reliability. Current scripts to annotate aligned phrases with probabilities go directly from relative counts to probabilities and do not apply thresholding. The information about absolute frequencies is lost, and it is rather hard to recover this info and apply it for improved weighting that takes reliability into account.

- If incorrect alignments or non-corresponding text segments are used to generate entries for SMT phrase tables, the problems typically do not affect isolated phrases, but regions within the training data where errors accumulate. Although single errors cannot be detected with high reliability, the fact that these errors arrive in bursts can be very helpful to identify problematic regions of the training data and to exclude them from the the data collection.

- Diversity in the data collection can also help to improve the quality. If a pair of corresponding phrases can be observed in many different contexts, e.g. in documents of different genres or from different domains, this correspondence is more reliable than if the same number of co-occurrences is observed in a single document or in a small number of very similar documents. If a correspondence can be found in resources of different types, e.g. both in parallel text and in a bilingual dictionary, it is again more reliable than if instances of it can be found only in resources of one specific type.

We have prepared an experimental setup to systematically exploit frequency information in the prediction of translation quality. This will be used in the upcoming evaluation campaign.

Another approach that can help to spot errors in parallel corpora, alignments, and phrase tables exploits the fact that correct correspondences can often be recovered indirectly, using additional intermediate languages. If some expression b is a translation of a in a intermediate language and c is a translation of b in the target language, then this is evidence[5] that c may be a correct translation of a. We can say that pairs (a, c) of expressions for which an expression b in an intermediate language can be found are more reliable than pairs for which this is not the case, and we can use multiple intermediate languages to obtain even stronger predictions for the reliability of correspondences in the translation model.

We have started to work on refined estimation of quality of phrase table entries based on intermediate languages and we will include the results of this work into the module for confidence estimation to be built within WP6.

[5]However, this is not true in all cases, as b may be ambiguous, and both instances of b may refer to incompatible meanings of b.

2.2.2 Improving recall

Whereas assigning non-zero probabilities to wrong entries in a phrase table is one important source of problems in statistical and hybrid MT, there is also the complementary problem that correct translations may be missing from the phrase table. This is typically due to the limited coverage of the examples found in the training data, aggravated by the very limited ability of current translations models to generalize existing examples to related cases, especially in richly inflected languages.

A number of research efforts have tried to address the problem of unseen words by integrating language-specific morphological information, allowing the SMT system to learn translations of base word forms. For example, Koehn and Knight (2003) showed how monolingual texts and parallel corpora could be used to figure out appropriate places to split German compounds. Niessen and Ney (2004) applied morphological analyzers to English and German and were able to reduce the amount of training data needed to reach a certain level of translation quality. Goldwater and McClosky (2005) found that stemming Czech and using lemmas improved the word-to-word correspondences when training Czech-English alignment models. de Gispert et al. (2005) substituted fully-inflected verb forms with their lemma to partially reduce the data sparseness problem associated with the many possible verb forms in Spanish. Kirchhoff et al. (2006) applied morpho-syntatic knowledge to re-score Spanish-English translations. Yang and Kirchhoff (2006) introduced a back-off model that allowed them to translate unseen German words through a procedure of compound splitting and stemming. Talbot and Osborne (2006) introduced a language-independent method for minimizing "lexical redundancy" by eliminating certain inflections used in one language which are not relevant when translating into another language. Talbot and Osborne showed improvements when their method is applied to Czech-English and Welsh-English translation. Other approaches have focused on ways of acquiring data in order to overcome problems with data sparsity. Resnik and Smith (2003) developed a method for gathering parallel corpora from the web. Oard et al. (2003) described various methods employed for quickly gathering resources to create a machine translation system for a language with no initial resources.

Callison-Burch (2007) takes a different approach to increase the coverage of translation models by introducing some amount of generalization into the training process through the use of paraphrases, which can be extracted automatically from bilingual parallel corpora.

These new ways of increasing the coverage of translation models trained on parallel data will not only help to find better translations for a given utterance, but will also help to make quality judgements on translations derived in different ways more meaningful. However, as most methods that try to increase the recall of models within a given framework, this new approach also runs a certain risk of "overshooting", i.e. including into the model pairs of expressions that are not mutual translations. The second half of the project will explore the potential of using automatically derived paraphrases and lexical entries found indirectly via a third language to derive better models of adequacy.

This will be done by augmenting the original phrase table with additional entries, which are distinguished from the ordinary (baseline) entries in an additional column for scores. We can then optimize models that include these entries and compare the results with the baseline setup. We can furthermore use the usual minimum error rate training

mechanism to optimize the weight of this extra column, which will allow us to fine-tune the penalty such additional entries should get in comparison to the baseline entries. Experiments along these lines will be done in connection to both upcoming evaluation campaigns (WMT08 and follow-up) and will be documented in the associated system descriptions.

Chapter 3

Architectures for Hybrid Machine Translation Systems

In this chapter, we motivate and give an overview of several different architectures that combine rule-based and statistical machine translation (RBMT and SMT) engines into hybrid systems. Two of these architectures have been implemented within the EuroMatrix project and will be investigated in some more detail in the upcoming evaluation campaigns. A third architecture is being developed in a cooperation with the European Patent Office, and feed-back from this institution, which makes active use of this setup in a real-world production scenario, will be extremely helpful for the further improvement of the method and for comparison with the alternatives.

3.1 Motivation of hybrid methods

Recent work on statistical and example-based MT has lead to significant progress, and has created viable alternatives to MT systems based on large rule systems.

However, for the translation into morphologically richer languages, where the well-formedness of an utterance in the target language cannot be judged using simple nGram statistics, systems based on SMT techniques alonge can only serve as a baseline implementation of MT functionality and need to be enriched with linguistic knowledge.

It turns out that the weaknesses and strengths of the different MT paradigms are somewhat complementarily distributed, as is sketched in Fig. 3.1.

Ideally, one would like to build up an integrated approach that combines the advantages of different paradigms, while at the same time avoiding the disadvantages. Such

	Syntax	Structual Semantics	Lexical Semantics	Lexical Adaptivity
Rule-based MT	++	+	-	–
Statistical MT	–	–	+	+
Example-based MT	-	–	-	++

Figure 3.1: Weaknesses and strengths of different MT paradigms

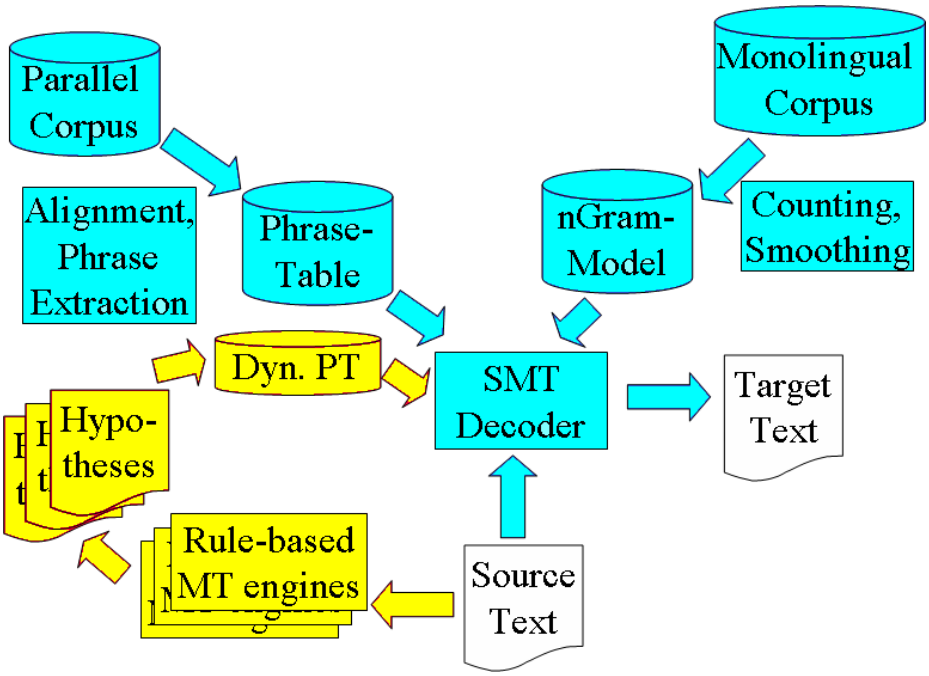

Figure 3.2: Architecture for multi-engine MT driven by a SMT decoder

an integrated approach should combine different types of knowledge in one architecture; explicit linguistic knowledge encoded in grammars, lexicons, and other rule systems with implicit knowledge about typical language usage[1] including weaker stylistic criteria.

The implementation of models that can encompass and integrate all theses types of knowledge from scratch is beyond the reach of a relatively small project that runs only for 30 months. Hence we were looking into ways to take as much existing technology as possible and to integrate it into a multi-engine architecture.

Combinations of MT systems into multi-engine architectures have a long tradition, starting perhaps with (Frederking and Nirenburg, 1994). Multi-engine systems can be roughly divided into simple architectures that try to select the best output from a number of systems, but leave the individual hypotheses as is (Tidhar and Küssner, 2000; Akiba et al., 2001; Callison-Burch and Flournoy, 2001; Akiba et al., 2002; Nomoto, 2004; Eisele, 2005) and more sophisticated setups that try to recombine the best parts from multiple hypotheses into a new utterance that can be better than the best of the given candidates, as described in (Rayner and Carter, 1997; Hogan and Frederking, 1998; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Matusov et al., 2006; Rosti et al., 2007).

Recombining multiple MT results requires finding the correspondences between alternative renderings of a source-language expression proposed by different MT systems. This is generally not straightforward, as different word order and errors in the output can make it hard to identify the alignment. Still, we assume that a good way to combine the various MT outcomes will need to involve word alignment between the MT output and the given source text, and hence a specialized module for word alignment is a central

[1]This also covers world knowledge, which is beyond the reach of today's formal models, but which can be approximated to some extent by observing statistical patterns in language use.

component of our setup.

Additionally, a recombination system needs a way to pick the best combination of alternative building blocks; and when judging the quality of a particular configuration, both the plausibility of the building blocks as such and their relation to the context need to be taken into account. The required optimization process is very similar to the search in a SMT decoder that looks for naturally sounding combinations of highly probable partial translations. Instead of implementing a special-purpose search procedure from scratch, we transform the information contained in the MT output into a form that is suitable as input for an existing SMT decoder. This has the additional advantage that it is simple to combine resources used in standard phrase-based SMT with the material extracted from the rule-based MT results; the optimal combination can essentially be reduced to the task of finding good relative weights for the various phrase table entries.

A sketch of the overall architecture is given in Fig. 3.2, where the blue (dark) parts represent the modules and data sets used in ordinary statistical MT, and the yellow (light) parts are the additional modules and data sets derived from the rule-based engines. It should be noted that this is certainly not the only way to combine systems. In particular, as this proposed setup gives the last word to the SMT decoder, we risk that linguistically well-formed constructs from one of the rule-based engines will be deteriorated in the final decoding step. Alternative architectures are under exploration and one such approach will be described below.

3.1.1 MT systems and other knowledge sources

For experiments in the framework of the shared task of the 2007 ACL workshop on SMT(Chen et al., 2007) we used a set of six rule-based MT engines that are partly available via web interfaces and partly installed locally. The web based systems are provided by Google (based on Systran for the relevant language pairs), SDL, and ProMT which all deliver significantly different output. Locally installed systems are OpenLogos, Lucy (a recent offspring of METAL), and Translate Pro by lingenio (only for German $\leftrightarrow$ English). In addition to these engines, we generated phrase tables from the training data following the baseline methodology given in the description of the shared task and using the scripts included in the Moses toolkit (Koehn et al., 2007). We enhanced the phrase tables with information on whether a given pair of phrases can also be derived via a third, intermediate language. We assume that this can be useful to distinguish different degrees of reliability, but due to lack of time for fine-tuning we could not yet show that it indeed helps in increasing the overall quality of the output.

3.1.2 Implementation Details

Alignment of MT output

The source text and the output text of the MT systems were aligned by means of GIZA++ (Och and Ney, 2003), a tool with which statistical models for alignment of parallel texts can be trained. Since training new models on merely short texts does not yield very accurate results, we applied a method where text can be aligned based on existing models

that have been trained on the Europarl Corpus (Koehn, 2005) beforehand. This was achieved by using a modified version of GIZA++ that is able to load given models.

The modified version of GIZA++ is embedded into a client-server setup. The user can send two corresponding files to the server, and specify two models for both translation directions from which alignments should be generated. After generating alignments in both directions (by running GIZA++ twice), the system also delivers a combination of these alignments which then serves as input to the following steps described below.

It should be noted that this way of computing the alignment between two short texts based on models trained on much larger corpora has other uses within a hybrid MT architecture. In particular, such a setup will help us to assign plausibility scores to the outputs of competing MT engines in a multi-engine setup, as sketched in Section 2.2. In case the quality of the alignments is not sufficient to derive meaningful estimates, it can be improved by training optimized alignment models for a given MT engine on data that is artificially generated using this engine.

Phrase tables from MT output

The standard phrase table (from the SMT baseline system) as well as all phrase tables obtained from the output of the rule-based MT systems were augmented by two additional columns, the first one indicating which MT system the phrase pair entry had been inherited from, the second column stating whether the phrase pair came from the standard phrase table (value 1) or from one of the rule-based MT systems (value 2). All of the phrase tables modified in this manner for a given translation direction were then concatenated and combined with "turned-around" versions of their respective counterparts, i.e. enhanced phrase tables for the opposite translation direction, thus forming one single large phrase table. The same procedure was also applied to all of the "reordering" phrase tables.

3.1.3 First Results

We compared the hybrid system to a purely statistical baseline system as well as two rule-based systems. The only differences between the baseline system and our hybrid system are the phrase table – the hybrid system includes more lexical entries than the baseline – and the weights obtained from minimum error rate training.

For a statistical system, lexical coverage becomes an obstacle – especially when the bilingual lexical entries are trained on documents from different domains. However, due to the distinct mechanisms used to generate these entries, rule-based systems and statistical systems usually differ in coverage. Our system managed to utilize lexical entries from various sources by integrating the phrase tables derived from rule-based systems into the phrase table trained on a large parallel corpus. Table 3.1 shows a rough estimation of the number of untranslated words in the respective output of different systems. The estimation was done by counting "words" (i.e. tokens excluding numbers and punctuations) that appear in both the source document and the outputs. Note that, as we are investigating translations from German to English, where the languages share a lot of vocabulary, e.g. named entities such as "USA", there are around 4.21% of words that should stay the same throughout the translation process. In the hybrid system, 5.59% of

Systems	Token #
Ref.	2091 (4.21%)
R-I	3886 (7.02%)
R-II	3508 (6.30%)
SMT	3976 (7.91%)
Hybrid	2425 (5.59%)

Table 3.1: Untranslated tokens (excl. numbers and punctuations) in output for news commentary task (de-en) from different systems

the words remain unchanged, which is is the lowest percentage among all systems. Our baseline system (SMT in Table 3.1), not comprising additional phrase tables, was the one to produce the highest number of such untranslated words.

	Baseline	Hybrid
test	18.07	21.39
nc-test	21.17	22.86

Table 3.2: Performance comparison (BLEU scores) between baseline and hybrid systems, on in-domain (test) and out-of-domain (nc-test) test data

Higher lexical coverage leads to better performance as can be seen in Table 3.2, which compares BLEU scores of the baseline and hybrid systems, both measured on in-domain and out-of-domain test data. Due to time constraints these numbers reflect results from using a single RBMT system (Lucy); using more systems would potentially further improve results.

3.2 Feeding SMT phrases into a rule-based MT system

3.2.1 Motivation

The architecture described in the last section places a strong emphasis on the statistical models and can be seen as a variant of SMT where lexical information from rule-based engines is used to increase lexical coverage.

However, also rule-based MT engines frequently suffer from missing lexical coverage, and it can be seen as a key advantage of SMT that lexical entries can be automatically induced from existing translations. It is therefore interesting to investigate how automatically extracted lexical knowledge can be used to increase the coverage of a rule-based MT system.

Such an arrangement leaves the control of the translation process with the rule-based engine, which has the advantage that well-formed syntactic structures generated via linguistic rules cannot be broken apart by the SMT components.

But as rule-based systems typically lack mechanisms for ruling out implausible results, they cannot easily cope with errors that creep into the lexicon due to misalignments,

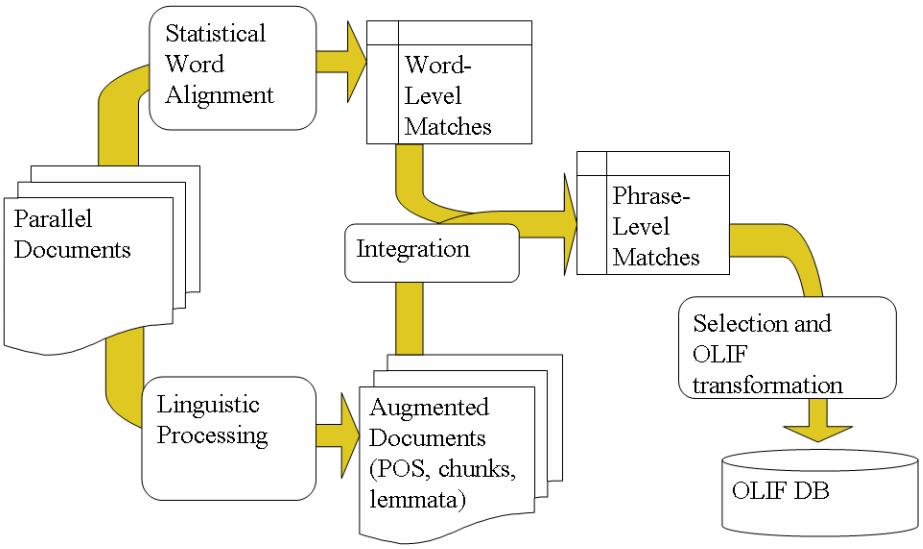

Figure 3.3: Bilingual terminology extraction to support rule-based MT

examples that fail to generalize, and similar problems.

Entries derived from statistical alignments need therefore to be carefully filtered to keep the error rate at an acceptable level. Furthermore, the information that can be extracted from word alignments of a given translation lacks linguistic information that is required by a rule-based system. Whereas corresponding expressions in a parallel corpus are found as inflected full forms, the entries in a bilingual dictionary contain normalized forms together with morphological classification that defines all possible inflectional forms of the given entry. Even if the parallel corpus happens to contain different forms of the entry, the collection of forms is a (typically very incomplete) random sample of the full paradigm from which it is not always possible to induce the complete inflectional behavior of the lemma.

Despite these additional difficulties, an infrastructure for the extraction of lexical entries was built up in the framework of a joint project between the DFKI and the European Patent Office (EPO), where the EPO wants to make translation functionality for patent documents available to their examiners and eventually also to the general public. The translation itself is done by an external service provider, using a rule-based MT engine, whereas the contribution of DFKI is the extraction and manual validation of additional lexical entries for the relevant technical fields.

In a selection test conducted in 2006, the EPO decided to base the MT functionality of this service on the technology provided by Systran.

3.2.2 Architecture for bilingual terminology extraction

Fig. 3.3 gives a schematic overview of the main modules used in this setup.

Parallel texts are on one hand sent through the statistical alignment machinery, based on GIZA++ that is also used for SMT to obtain word and phrase alignments. On the other hand the texts are linguistically enriched by part of speech (PoS) tags and

lemma information. The two representations are then combined and filters based on PoS sequences on both sides are used to obtain a set of candidates for the lexicon. A list of acceptable pairs of PoS sequences is generated by inspecting several hundred of the most frequently occurring PoS sequences and excluding those that either do not form a pair of linguistic phrases or where the interpretation on both sides is incompatible. Morphological classification is applied to these lexical entries to augment them with inflection classes, following the open lexicon interchange format (OLIF) standard (Lieske et al., 2001).

In a first round of extraction work about 40 million English-German sentence pairs and about 10 million English-Spanish sentence pairs have been processed and 2.3 million candidates for English-German term pairs as well as 0.8 million candidates for English-Spanish term pairs have been identified. About 90% of the extracted entries are pairs of noun phrases, which typically consist or multi-word expressions (MWEs) involving one or more adjectives or noun compounds[2].

An evaluation by the EPO showed that a significant subset of the identified term pairs are either correct or close enough to correct lexical entries that manual validation or correction seems worthwhile.

3.2.3 Infrastructure for terminology validation

Even if statistical alignment and linguistic preprocessing can lead us a long way towards the automatic creation of lexical entries, it is crucial to manually inspect, filter, and correct the resulting candidates, as a rule-based MT systems offers no other mechanism to prevent errors caused by wrong lexical entries. In cases of technical terminology, the validation of the terminology requires both linguistic and technical competence, and it may be necessary to distribute some steps over different groups of people.

In order to facilitate this process, we have built a web-based front end for lexical database maintenance such that the extracted lexical entries are stored in a centralized way and various parts of validation and quality control can be distributed over arbitrary workplaces that have access to the internet.

The validation workflow consists of several steps where the entries are first checked for monolingual linguistic wellformedness and classified according to morphologic properties like head, gender, and inflectional class. This part of the interface is built such that the validator does not see internal codes for the inflectional class, but sees a small set of distinctive full forms and has an option to correct these[3].

In a second round the corresponding forms from two languages are seen in combination and the validator can rule out the cases where the forms do not convey the same meaning. This round also deals with disambiguation; whereas generally for a term in one language the most frequently appearing translation in the same subject domain is used,

[2]Often, English MWE correspond to one long German word, like Kathodenstrahlröhrensteuerungsanordnung = CRT controller, Hydroxypropylmethylcelluloseacetatsuccinat = hydroxypropylmethylcellulose acetate succinate, Empfängnisverhütungsmittelzusammensetzung = contraceptive composition, Unempfindlichkeitsbereicherzeugungsschaltkreis = deadband generating circuit, Datenübertragungsblocksynchronisationserfassungseinrichtung = frame synchronization detector

[3]For German nouns it is sufficient to check the singular forms for Nominative and Genitive and (if possible) the plural forms for Nominative and Dative

the validator has the choice to disprefer certain expressions. Dispreferred expressions will then still be understood in the source text, but will be avoided in the target text in favor of expressions that appeared less frequently.

In a third round the DB interface is used by representatives of the participating patent authorities for quality control by domain experts.

3.2.4 Results

The proposed architecture was used to create translation dictionaries with technical vocabulary for all four language directions (EN paired with ES or DE, in both directions).

For each direction, 60000 lexical entries were selected by the EPO and manually validated by linguists at DFKI. As the entries are derived from documents for which the technical domain is known, it is possible to annotate the entries with the frequencies in which this translation is encountered in documents from this particular domain[4].

Using this simple mechanism, it is possible to use knowledge of the IPC class of the document to be translated to select the most appropriate translation of a given term in the source language.

Comparisons of the translation quality with and without the automatically derived translation entries revealed a significant increase in lexical coverage using our model.

3.3 Other Architecures for Hybrid MT

The two architectures discussed in the preceeding sections are in a certain way complementary. The architecture sketched in Fig. 3.2 is essentially an extension of phrase-based SMT by a mechanism to acquire additional lexical entries from the lexicon of one or several rule-based MT engines. The architecture in Fig. 3.3 is an extension of a rule-based MT engine by a mechanism to learn lexical entries from parallel corpora. Both architectures share the advantage that acquisition of lexical data is facilitated compared to a "pure" incarnation of the respective paradigm. However, both architectures also share the disadvantage that integration of multiple knowledge sources is not as flexible as it should be. In the SMT case, there is no way to make use of syntactic or morphologic constraints, so we cannot expect a significant reduction in the types of grammatical errrors that are typical for state-of-the-art SMT. The rule-based architecture does not contain new mechanisms that would take soft constraints and preferences into account, and the granularity of the lexical units is limited to the types that are supported by the underlying rule mechanism. We should hence not expect dramatic improvements in the overal fluency of translations.

These considerations lead to the question how a tighter integration of knowledge sources could be implemented that would give different types of knowledge a more balanced weight. Preferrably, such a more tightly integrated architecture would still make use of the main building blocks used in existing MT architectures, but

In (Chen et al., 2007), we have proposed a couple of additional architectures that could help to alleviate these problems.

[4]We use the international patent classifiation (IPC) for these distinctions, see http://www.wipo.int/classifications/fulltext/new_ipc/ipcen.html

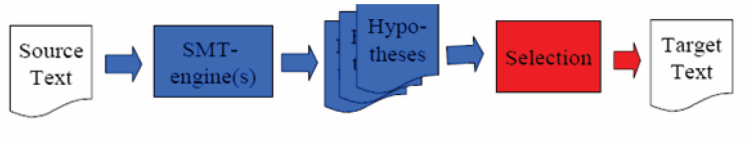

Figure 3.4: Using syntactic constraints in the selection of MT output

3.4 Using Syntactic Constraints in a SMT architecture

As one of the main problems of today's SMT models is the lack of grammatical knowledge that is taken into account, it looks natural to make use of such constraints when selecting the best out of a large set of hypotheses considered by a SMT decoder.

In its purest form, this consideration would then lead to an architecture as depicted in Fig. 3.4, which is similar to the basic SMT setup, except for the fact that selection does not happen on the basis of statistical LM scores, but on the basis of checks of syntactic constraints.

It should be clear that replacing statistical LM scores completely by syntactic constraints would be too radical. It would ignore important knowledge contained in nGram statistics and would also lead to a couple of practical problems with robustness and efficiency of the grammars and parsers used for testing syntactic well-formedness. Instead, different types of knowledge need to be combined into a consolidated estimate of quality, so that an intelligent combination of this architecture with the baseline SMT setup will in the end achieve better accuracy than the nGram based language models.

(Och et al., 2004) describes how a multitude of features and constraints, including shallow syntactic ones, can be taken into account in the selection of a translations, and they give an extensive comparison of many SMT models composed from various knowledge sources. Whereas the improvements reported in this work do not appear dramatic, this could also be due to the fact that the chosen language pair (Chinese to English) does not display the full potential of the approach. In order to broaden the scope of our investigation into hybrid MT architectures, we plan to perform similar experiments for language pairs that involve a richer set of morphological constraints and to compare them with baseline performance based on simple nGram models.

3.5 Interleaving grammar-based transfer with stochastic ranking

One of the most important problems with purely rule-based MT architectures is the difficulty to distinguish between outcomes that are all licensed by the given set of rules, but which appear more or less plausible based on additional knowledge that cannot be captured in hard rules.

(Oepen et al., 2007) present the results of the Logon project and discuss ways how a strongly rule-driven architecture based on grammars for source and target language as

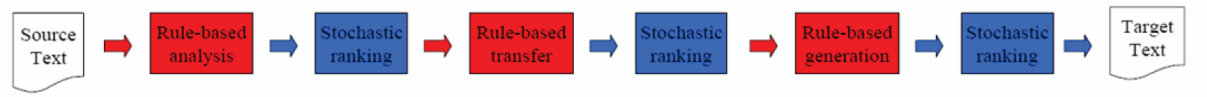

Figure 3.5: Integrated usage of rule-based and stochastic knowledge in a transfer architecture

well as on a system of transfer rules can be enriched with statistical knowledge sources and preferences derived from corpora.

This leads to an architecture as sketched in Fig. 3.5, where rule-based generation of (intermediate) representations alternate with stochastic selection of smaller sets of candidates to be used in subsequent steps. From all the architecures presented so far, this setup seems to provide the best integration of multiple knowledge sources. However, the full potential of this architecture can only be shown if the grammars on both sides as well as the system of transfer rules achieve high coverage and furthermore if the representations used in the three rule systems are compatible, so that no intermediate results are lost due to structural mismatches. (Jellinghaus, 2007) presents methods to learn the transfer rules required in such an architecture from pairs of structures derived from a parallel corpus. These techniques, which are also described in EuroMatrix Deliverable 4.1, will make experiments with this architecture much easier than if the transfer rules would have to be written by hand, as was the case in the Logon-Project.

3.6 Statistical post-correction of rule-based MT output

Researchers both from within the EuroMatrix consortium (Dugast et al., 2007) and from other research teams (Simard et al., 2007) have recently proposed a combination of rule-based MT with statistical techniques where the main function of the statistical MT engine is the correction of the output of the rule-based engine.

The main idea of this approach is to use a parallel corpus as a set of training data not for the mapping from one language to the other, but to map the output of an MT system into a version that hopefully avoids many of the typical errors of the MT system.

In order to investigate the quality of this combination and to compare typical errors of this setup with some of the alternatives proposed above, (Theison, 2007) implemented both the post-correction method and the approach of Fig. 3.2 and did a careful investigation of the various types of errors found in both cases.

This comparison revealed that, whereas the architecture of Fig. 3.2 shows typical errors of an SMT system, the number of such SMT-typical errors is reduced in the post-correction architecture. On the other hand, many of the typical errors of rule-based MT engines cannot be corrected using SMT-based post-processing, resulting in a rather similar total number of errors.

So far, we have not yet found an effective way to achieve a significant reduction of errors of one type without at the same time increasing the frequency of other error types.

Chapter 4

Conclusion and Outlook

In Chapter 3 of this document, we have presented two complementary ways to combine rule-based and statistical approaches to MT by integrating existing implementations into a larger architecture. In one case, rule-based MT engines are used to enrich the lexical ressources available to the SMT decoder, in the other case parts of the SMT infrastructure are used, together with linguistic processing and manual validation, to extend the lexicon of a rule-based MT engine. Both approaches have been implemented and show promising improvements to MT quality[1], but as they are currently still in a somewhat prototypical state, it is still too early to give meaningful comparative evaluations.

However, it should be clear that even if one or both of these approaches can be made to deliver significant improvements under fairly general conditions, the improvements will essentially only alleviate the problem of lexical coverage, but will not touch some other well-known issues with the respective frameworks. One of the key problems of rule-based MT systems is their difficulty to deal with soft rules and preferences that are required for disambiguation and for picking the most natural expressions in the target language. Conversely, today's versions of SMT have obvious difficulties delivering syntactically well-formed utterances, especially when relevant contraints reach beyond the window size of the target language models. It is conceivable that a deeper integration of rule-based linguistic knowledge with corpus-based evidence will eventually be able to alleviate both problems in one integrated system, but this will require an architecture that has simultaneous access to all relevant type of knowledge, which is beyond the relatively simple hybrid architectures presented here.

[1]Error statistics concerning the lexically extended SMT setup are given in (Theison, 2007). In the case of the extended rule-based MT, we currently need to rely on the feed-back by the European Patent Office, as we do not have direct access to the augmented MT installation

References

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Yasuhiro Akiba, Taro Watanabe, and Eiichiro Sumita. 2002. Using language and translation models to select the best among outputs from multiple mt systems. In *COLING*.

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *ASRU*, Italy.

Chris Callison-Burch and Raymond S. Flournoy. 2001. A program for automatically selecting the best output from multiple machine translation engines. In *Proc. of MT Summit VIII*, Santiago de Compostela, Spain.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June. Association for Computational Linguistics.

Chris Callison-Burch. 2007. *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, November.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source SMT decoder. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 193–196, Prague, Czech Republic, June. Association for Computational Linguistics.

A. de Gispert, J.B. Mariño, and J.M. Crego. 2005. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Proc. of the 9th European Conf. on Speech Communication and Technology (Interspeech'05)*, Lisboa (Portugal), September.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223, Prague, Czech Republic, June. Association for Computational Linguistics.

Andreas Eisele. 2005. First steps towards multi-engine machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 155–158, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Marcello Federico and Mauro Cettolo. 2007. Efficient handling of n-gram language models for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 88–95, Prague, Czech Republic, June. Association for Computational Linguistics.

Christian Federmann. 2007. Very large language models for machine translation. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany, July.

Robert E. Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *ANLP*, pages 95–100.

Sharon Goldwater and David McClosky. 2005. Improving statistical mt through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Christopher Hogan and Robert E. Frederking. 1998. An evaluation of the multi-engine mt architecture. In *Proceedings of AMTA*, pages 113–123.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. of EAMT*, Budapest, Hungary.

Michael Jellinghaus. 2007. Automatic Acquisition of Semantic Transfer Rules for Machine Translation. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany.

K. Kirchhoff, M. Yang, and K. Duh. 2006. Statistical machine translation of parliamentary proceedings using morpho-syntactic knowledge. In *Proc. of the TC-STAR Workshop on Speech to Speech Translation*, Barcelona, Spain.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of EACL'03: 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 00–00, Budapest, Hungary.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X.*

Christian Lieske, Susan McCormick, and Gregor Thurmair. 2001. The open lexicon interchange format (olif) comes of age. In *Proceedings of MT Summit VIII: Machine Translation in the Information Age*, pages 211–216, Santiago de Compostela, Spain, September.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *In Proc. EACL*, pages 33–40.

Sonja Niessen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

Tadashi Nomoto. 2004. Multi-engine machine translation with voted language model. In *Proc. of ACL.*

Douglas W. Oard, David S. Doermann, Bonnie J. Dorr, Daqing He, Philip Resnik, Amy Weinberg, William J. Byrne, Sanjeev Khudanpur, David Yarowsky, Anton Leuski,

Philipp Koehn, and Kevin Knight. 2003. Desparately seeking cebuano. In *HLT-NAACL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of NAACL-04*, Boston.

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards hybrid quality-oriented machine translation — on linguistics and probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, Skövde, Sweden.

Manny Rayner and David M. Carter. 1997. Hybrid language processing in the spoken language translator. In *Proc. ICASSP '97*, pages 107–110, Munich, Germany.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007. Combining translations from multiple machine translation systems. In *Proceedings of the Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL'2007)*, pages 228–235, Rochester, NY, April 22-27.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206, Prague, Czech Republic, June. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*.

David Talbot and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of ACL*.

Silke Theison. 2007. Optimizing Rule-Based Machine Translation Output with the Help of Statistical Methods. Diploma thesis, Universität des Saarlandes, Saarbrücken, Germany.

Dan Tidhar and Uwe Küssner. 2000. Learning to select a good translation. In *COLING*, pages 843–849.

M. Yang and K. Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proc. of the EACL*.

Yi Zhang. 2007. *Robust deep linguistic processing*. Ph.D. thesis, Universität des Saarlandes, Saarbrücken, Germany, December.