Chapter 6
Getting Past the Language Gap: Innovations in Machine Translation

Rodolfo Delmonte

Abstract In this chapter, we will be reviewing state of the art machine translation systems, and will discuss innovative methods for machine translation, highlighting the most promising techniques and applications. Machine translation (MT) has benefited from a revitalization in the last 10 years or so, after a period of relatively slow activity. In 2005 the field received a jumpstart when a powerful complete experimental package for building MT systems from scratch became freely available as a result of the unified efforts of the MOSES international consortium. Around the same time, hierarchical methods had been introduced by Chinese researchers, which allowed the introduction and use of syntactic information in translation modeling. Furthermore, the advances in the related field of computational linguistics, making off-the-shelf taggers and parsers readily available, helped give MT an additional boost. Yet there is still more progress to be made. For example, MT will be enhanced greatly when both syntax and semantics are on board: this still presents a major challenge though many advanced research groups are currently pursuing ways to meet this challenge head-on. The next generation of MT will consist of a collection of hybrid systems. It also augurs well for the mobile environment, as we look forward to more advanced and improved technologies that enable the working of Speech-To-Speech machine translation on hand-held devices, i.e. speech recognition and speech synthesis. We review all of these developments and point out in the final section some of the most promising research avenues for the future of MT.

R. Delmonte, Ph.D. (✉)
Department of Linguistic Studies and Comparative Cultures, Ca' Foscari University, Dorsoduro 1075, Venezia 30123, Italy
e-mail: delmont@unive.it; project.cgm.unive.it

A. Neustein and J.A. Markowitz (eds.), *Mobile Speech and Advanced Natural Language Solutions*, DOI 10.1007/978-1-4614-6018-3_6,

Introduction

In 2005b John Hutchins wrote the following gloomy assessment of Machine Translation (MT):

> Machine translation (MT) is still better known for its failures than for its successes. It continues to labour under misconceptions and prejudices from the ALPAC report of more than thirty years ago, and now it has to contend with widespread misunderstanding and ridicule from users of online MT services. The goal of developing fully automatic general-purpose systems capable of near-human translation quality has been long abandoned. The aim is now to produce aids and tools for professional and non-professional translation which exploit the potentials of computers to support human skills and intelligence, or which provide rough translations for users to extract the essential information from texts in foreign languages. JH (ibid., 1–5)

Since then the field of Machine Translation (MT) has dramatically changed. And in the past 3 years, the field of MT has become so huge that there is no chance of sufficiently reviewing the whole spectrum of activities, tools and resources related to the field. Therefore, I will restrict this discussion to the leading and most promising approaches.[1] The first section will be devoted to examining in detail what Statistical Machine Translation (SMT) can offer in terms of improvements to the state of the art. Then, I will dedicate section "Hybrid and Rule-Based MT Systems" to hybrid methods and systems. In section "Syntax Based Approaches: From Hierarchical to SBSMT", I will delve into syntactically based MT systems. Then section "Knowledge-Based MT Systems" will introduce knowledge, semantically-based systems. Section "Evaluation Methods and Tools" will comprise an overview of evaluation methods and section "MT for the Future" will briefly give an overview of what in my opinion may constitute the future of MT systems. This section will comprise a subsection dedicated to Speech-To-Speech MT; in another subsection promising national projects will also be reviewed. This followed by the last section in which I draw conclusions.

As JH noted, SMT research now dominates MT research. In spite of that, the great majority of commercial systems are Rule Based MT (RBMT) systems. Also most if not all professional translators are not using any of the research products (pp. 1–5). SMT systems that have reached public operational status are still only few in number, and perhaps "LanguageWeaver" – an offshoot of the research group at the University of Southern California – can be regarded as the best system offering translation systems for Arabic, Chinese, and most European languages to and from English. Always quoting from JH (ibid., 5–7):

> … there is great and increasing usage of web-based MT services (many free), such as the well-known 'Babelfish' available on Yahoo. Others include FreeTranslation, Google Translator, Bing Translator, Tarjim, WorldLingo.

[1] Consequently, I will not be concerned with commercial versions of MT systems, nor to the field of computer-assisted Translation systems or translation aids: they are all listed at http://www.hutchinsweb.me.uk/Compendium.htm, a document produced almost on a yearly basis and compiled by John Hutchins, who is also responsible for the main source of information on MT which is regularly posted on the Machine Translation Archive (http://www.mt-archive.info).

> … there are three systems specifically for translating patents: the PaTrans and SpaTrans systems developed for LingTech A/S to translate English patents into Danish …
>
> Online services are now predominantly SMT-based, e.g. 'Google Translate', 'Bing Translate' (previously 'Windows Live Translator'), 'Babelfish' (now on the Yahoo site).
>
> Probably the most significant development in MT research in Europe is the establishment of the Euromatrix project (based at Edinburgh University). Its aim is the development of open-source MT technologies applicable to all language pairs within Europe, based on hybrid designs combining statistical and rule-based methods. Perhaps best known is the Apertium framework, used for systems for Spanish, Catalan, Portuguese and Basque.

As JH noted, and I also believe, hybridization is the most interesting development of MT, and a section below will be devoted to careful examination of hybrid systems and methods. RBMT systems have been combined with SMT, and multiple subsystems are used in conjunction, such as morphological analysers, dependency parsers, and semantic engines in combination with Phrase-Based MT (PBMT). On the other side, hybrid systems that take advantage of examples have come to be used thanks to the availability of big parallel corpora of examples or translation memories, such as DGT-TM, a translation memory (sentences and their manually produced translations) organized by Ralf Steinberger from JRC and taken from the corpus Acquis Communautaire, freely downloadable at http://langtech.jrc.ec.europa.eu/DGT-TM.html. It contains all 231 language pairs from the European 22 languages, for a total of about three million sentences for most languages, 57 million in total.

Languages covered by MT have now dramatically increased, covering all European language pairs. But also Arabic and East Asian languages have become commonly translatable by MT tools, including Korean, Chinese, Japanese, Tahi, Urdu, Vietnamese, Bengali, Punjabi, Hindi, etc. Of particular interest to the US government bodies are languages like Pashto and Farsi which have also been object of translation engines.

The multi-engine approach involves the translation of a given text by two or more different MT architectures (SMT and RBMT, for example) and the integration of outputs for the selection of the 'best' output – for which statistical techniques can be used. The idea is attractive and quality improvements have been achieved, but it is difficult to see this approach as a feasible economical method for large-scale or commercial MT. An example of appending statistical techniques to rule-based MT is the experiment (by a number of researchers in Spain, Japan, and Canada) of 'statistical post-editing'. In essence, the method involves the submission (for correction and improvement) of the output of an RBMT system to a 'language model' of the kind found in SMT systems. One advantage of the approach is that the deficiencies of RBMT for less-resourced languages may be overcome. There will be more discussion on this topic below.

Statistical MT: Strength and Weaknesses

In SMT the task of translating one sentence from a source into a target language is transformed into the task of finding the "best" translation with minimum error rate: this is technically also called the minimum loss decision. In order to compose a

complete sentence, machine translation systems score small units of translation and select the fragments that, when combined together, yield the best score according to their model. The basic components of a SMT are three: a translation model, a language model and a decoder. Phrase-based SMT works as follows: source input is segmented in phrases (any sequence of words); each source phrase is automatically aligned to a target phrase on the basis of word alignment; and, eventually phrases are reordered. The decoder is responsible for the choice of best translation at sentence level: it builds translation monotonically from left to right, and the other way around for languages like Arabic and Chinese. It collects all phrase pairs that are consistent with word alignment and finds the best candidate phrase. Then it adds it to the end of partial translation, at the same time it marks the source phrase as translated. At the end of the decoding process there may be reordering. Phrase translation is the core process. There are many possible ways of segmenting and translating phrases: this is done on a probabilistic basis, and the probability distribution of the collected phrase pairs is usually based on their relative frequency. The task could be then rephrased as finding the best translation candidate hypothesis that covers all words/phrases in a sentence. Weak hypotheses are discarded and the best path is the one with best candidates. At each step the algorithm estimates costs to translate remaining part of input, and tries to find the cheapest sequence of translation option for each adjacent span of text.

Statistical MT research has explored the use of simple phrases (Och and Ney 2004), Hiero grammars (Chiang 2005), and complex S-CFG rules (Zollmann and Venugopal 2006). These more specialized translation units can more accurately describe the translation process, but they are also less likely to occur in the corpus. The increased data sparsity makes it difficult to estimate the standard SMT features which are typically computed as relative frequencies. Current weaknesses and permanent flaws are:

- Wrong word choices
- Presence of unknown words or OOVWs (Out Of Vocabulary Words)
- Mangled grammar
- Difficulties in treatment of function words (locally adding, dropping, changing)
- Lack of syntactic transformations for long-distance dependencies which require some reordering
- Lack of translation consistency (as argued by Xiao et al. 2011)
- Lack of resiliency in presence of morphologically rich languages (Chinese and English are better suited just because they are morphologically poor)
- Lack of sufficient contextual information both in translation model and in language model (trigrams are insufficient to model language discontinuities – however Chinese-English STM use a 5-gram language modeling, Shujie Liu et al. 2011), but see below.

In addition to phrase translation models, also word translation models are used, based on lexical level translation in conjunction with PBSTM. Word-based MT has a number of deficiencies that have been considered when moving to phrase-based MT, which, however, are worth while commenting.

– IBM models used have the possibility of matching one source word to many output target word – this is called fertility of the model – but not the opposite;
– Word-based models miss the majority of collocations and multiword expressions
– For many languages the syntactic structure is not symmetric and requires reordering: however word-based models penalize any such reordering and are not capable of enforcing the positioning of words at totally different places in the sentence – say the verb in Chinese and Japanese at the end of the sentence or in Arabic at the beginning compared to SVO languages.

In phrase-based models, on the contrary, the more data are available, the longer the phrases are learned, and in some cases whole sentences can be learned. Thus local context can be taken into consideration fully – the only problems may come from syntactic discontinuities and long distance dependencies, as indicated above.

Alignment of phrases goes in both directions and in this case allows for optimized results – always with IBM3 model. After aligning in both directions the results are merged and the best union or intersection is kept. As said above, phrase alignment must be consistent with word alignment and cover all words in the sentence. In this way, phrases are induced from words level alignments. Probabilities for phrases are just relative probabilities associated with each word in the phrase – a summation or a multiplication of them:

– Probability(Source/Target-phrase)=count(Source/Target-phrase)/count(Target-phrase)

The Problem of Word Alignment

Alignments are produced on the basis of IBM models 1–5, which I briefly review here (but see also Koehn 2010). Model 1 assumes that given a certain sentence length, the possible connections of words from Source to Target are all equally possible – in this way their actual order has no impact. Model 2 introduces probability to the connection between words in S/T, and it is based on position and length of the string. Models 3, considers the number of possible connections from S to T in a many-to-one fashion – thus allowing missing words or fertility. This is further conditioned in Model 4 by the identity of the words to connect in S/T. Model 5 is used to fix deficiencies. So eventually, Model 1 does lexical translation, Model 2 adds some reordering model to the output of Model 1. Then in Model 3 a fertility model is added and in Model 4 a relative reordering model, or distortion model is created always on a probabilistic basis. Distortion models are necessary every time one-to-one or monotonic alignment is insufficient, and usually ensues from many-to-many mappings (see Tiedemann 2011). The many-to-one fertility translation model is exemplified best with reference to an English to German translation system, where compound German words have to be aligned to many English words. But in some cases, English may have phrasal expressions instead of a single word in German,

so the opposite is needed. So more generally, many-to-many alignments are needed and this can be done by using GIZA++ (Och and Ney 2004) bidirectionally. The translation model is computed on the basis of word alignment, and this is regarded a critical component in SMT. However word alignment is automatically induced from parallel data, and this is usually what may constitute a real bottleneck, at least as happens in not related language pairs, where accuracy is below 80% (Hermjakob 2009). In order to produce a complete word alignment at sentence level, the system passes through the text for up to 20 iterations, to find frequent co-occurrences of words, that is words occurring in the same position in both source and target text. This usually happens at the beginning with most frequent words, that is function words – articles, prepositions, conjunctions, etc. – less frequently for content words which are more sparse. Thus eventually probabilities for content words may easily go up to 1.0, if suppose a word like "book" cooccurs with article "the" all the time. The system will look for alignments in adjacency of an already aligned word in case of misalignments: some words may come before or after another word depending on language grammar. Typically this applies to adjectives in English and Italian for instance, where English has the majority before the head noun and Italian after the head noun. Phrases will typically cover all local linguistically related positional differences in word order: decoding or translating is done monotonically once phrase alignment is terminated. Different types of constraints are applied to alignment processes as regards for instance the maximum size of segments involved in the mapping; or the maximum distance allowed for aligned segments with respect to their position in a distortion model. Translation modeling as presented above, comprises three steps: at first, sentence pairs in training corpus are aligned at word level. Then, translation pairs are extracted using some heuristic method. Lastly, maximum-likelihood estimation (MLE) is used to compute translation probabilities. The most relevant shortcoming of this method is possible inconsistent format of translation knowledge: word alignment in training versus translation pairs (phrase pairs) in decoding; then the training process which is not oriented towards translation evaluation metric (BLEU not being considered in the scoring of translation pairs – but only error minimization procedures). In this way, it is not possible to know whether translation phrases are extracted from a highly probable phrase alignment or from an unlikely one. In fact the incorrect phrases induced from inaccurate word-aligned parallel data is one of the major reasons for translation errors in phrase-based SMT systems. In Fig. 6.1 below I show the pipeline of a generic SMT using Moses (Moran et al. 2007).

Learning Improves Performance

Learning has been applied to the final decoding phase by introducing weights associated with translations and a final phase in which automatic evaluation is applied to the output of the system. This has improved dramatically the performance of SMT (see Saers et al. 2010; Saers 2011). Learning in this case is just finding model weights that make the correct translations score the best: to this aim procedures and techniques are directed to creating an optimizer, as discussed below (but see also Ambati et al. 2011).

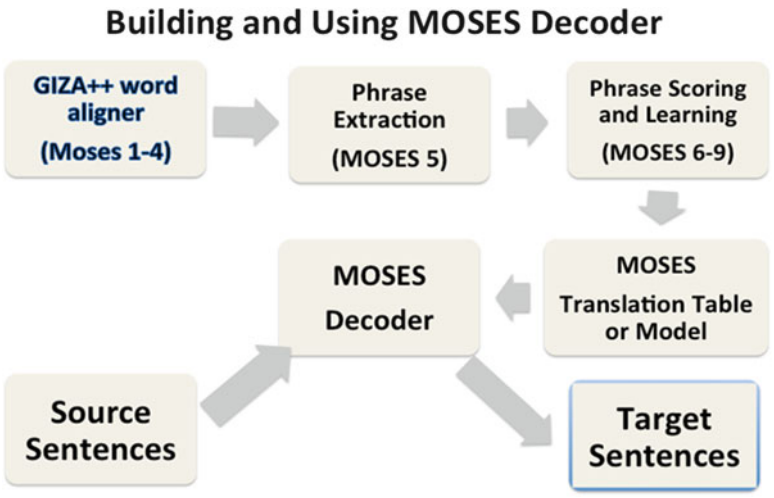

Fig. 6.1 A generic typical pipeline for an SMT system using MOSES

Discriminative models have been introduced so that translations are ranked and learned automatically by the use of features. A model consists of a number of features (e.g. the language model score). Each feature has a weight by measuring its value for judging a translation as correct. Then feature weights are optimized on training data, so that the system output matches correct translations as closely as possible. Feature weights can be adjusted and the process iterated a number of times – typically 20 iterations. Learning weights is done in a loop where the decoder generates the n-best list of candidate translation pairs. These are scored by an automatic evaluation tool – typically BLEU – then a reranking takes place which allows the system to learn best features that qualify best translations. This allows the system to change feature weights. Searching for the optimal parameters in linear models (Och and Ney 2002) of SMT has been a major challenge to the MT community. Statistical methods try to improve translation quality by minimizing error rate, and the most widely used approach to-date is Minimum-Error-Rate Training (MERT:Och 2003), which tries to find the parameters that optimize the translation quality of the one-best translation candidate, using the N-best list as an approximation of the decoder's search space. In this way, the system tries to find the best parameter that optimizes the translation quality of the first best translation candidate, and reranking follows. Reranking is done on the basis of MERT, however this method is unstable. As Cettolo et al. (2011) observe, in the last years, many efforts have been devoted to making the decoding procedure or its results more reliable. Recently, a deep investigation of the optimizer instability has been presented by Clark et al. (2011). Statistical machine translation (SMT) systems are based on a log-linear interpolation of models. Interpolation weights are typically computed by means of iterative procedures which aim at optimizing the scores of a given function. Unfortunately, as Cettolo et al. (2011) note, such a function is definitely non-convex; hence, only local optima can be reached. Moreover, it has been observed that the commonly used optimization

procedure, the N-best MERT is quite unstable. Now the focus is on improvement by reducing error rate as measured by evaluation methods. As Phillips (2011) comments, in spite of its usefulness and high adoption, MERT suffers from shortcomings of which the MT community is becoming aware. On the one hand, MERT is not designed for models with rich features and therefore leads to translations of unstable quality in such scenarios. The fluctuation in quality can even be statistically perceivable when the number of features is larger than 25 or 30 in practice. On the other hand, Smith (2006) finds that, MERT relies heavily on the behavior of parameters on the error surface, which is likely to be affected by random variances in the N-best list, and also lead to less generalizable results especially when the development set and the test set are not from exactly the same domain. As Phillips (2011) remarks, a significant challenge in building data-driven MT systems is identifying the right level of abstraction–to model translation units that both adequately reflect the data and can be estimated well.

System combination is another technique to rank the best translation, which has been applied extensively to SMT. One research line takes n-best translations of single systems, and produces the final output by means of either sentence-level combination, i.e. a direct selection from original outputs of single SMT systems (Sim et al. 2007, Hildebrand and Vogel 2008), or phrase- or word-level combination, i.e. the synthesis of a (possibly) new output joining portions of the original outputs (Rosti et al. 2007a, b, 2008; Ayan et al. 2008; He et al. 2008). These works focus on the combination of multiple machine translation systems based on different models and paradigms. More on these proposals in the section below.

Translation Models and the Problem of Overfitting

It is possible to distinguish between generative translation models (essentially, the IBM models), and the other half to various discriminative models. The first type of models induce a full probability distribution including both target and observable data and work in an unsupervised manner. The second type, on the contrary, work on labeled training data, thus supervised or semi supervised and suffer from usual related problems like data sparsity (see Tiedemann 2011:pp. 17–18). Ravi and Knight (2010) after experimenting with GIZA (sub-optimal hill-climbing) Viterbi alignment and comparing it to optimal version cast in integer linear programming (ILP), have determined that GIZA++ makes few search errors (between 5% and 15%), despite the heuristic nature of the algorithm, and that these search errors do not materially affect overall alignment (F-measure) accuracy, seen that Chinese-English averages 57–65%, and Arabic-English at 43–55% – with best values for the version that has English as target. Now, as indicated above, words that occur in totally different sentence positions, or function words that don't occur in some languages may result in poor word/phrase alignment. This problem is discussed particularly in Ulf Hermjakob's (2009), where he suggests the use of linguistic knowledge, in particular syntactic parse trees and the use of gazetteers for named-entity recognition and amalgamation. More on this topic below.

The problem of phrase-pair creation based on word alignment is nicely approached in the paper by Hyoung-Gyu Lee et al. (2010). The authors take into consideration the "collocation"-like ability of adjacent words to appear in a phrase. Different phrase segmentation will generate different translation results. To prevent bad phrases from being assigned high probabilities both collocation properties of words and multiword related probabilities taken from a corpus may be important to use. The authors characterize conditions of good segmentation which is necessary to produce high quality translation. The segmentation model they propose will consider lexical cohesion of adjacent words and the translational diversity of a word sequence as characteristics of good segmentation. To associate probabilities to such notions, they use collocation statistics from a corpus and translational entropy measures (see also Liu et al. 2010). The second is exemplified as follows: a phrase that has high translational entropy at word level but whose translational diversity at phrase level is low, should not be segmented. Though individual words in a phrase may be diversely translated with a high number of different translation pairs, the phrase may be translated with only few expressions. This would be typical of idiomatic expressions, and their model will score them very high.

Translation models should have a double function. They should well represent the training data – no gap and no bad translation; at the same time they should be able to generalize to unseen datasets.

Shujie Liu et al. (2011) discuss the problem of overfitting and note that during the training phase, the possibility of overfitting is always present. Consequently, this will hamper generalizing to unseen data: training should always be accompanied by a test phase with different datasets from the training ones. But this may not be sufficient. In fact, as the authors comment, the training phase is itself questionable because it usually optimizes on the feature weights associated to each sub-model (translation, fertility, distortion, etc.) rather than on the phrase-based translation model. At the same time, PBMT creates the phrase-based translation pairs on the basis of word alignment and the probabilities assigned by maximum-likelihood estimation (MLE). The paper proposes a new unified framework to add a discriminative translation sub-model into the conventional linear framework (more on discriminative models below), and the sub-model is optimized with the same criterion as the translation output is evaluated (BLEU in our case). In this case, each translation pair is a feature in itself, and the training method can affect the pairs directly so as to handle overfitting (ibid., 181).

As the authors clearly demonstrate, the translation model will overestimate probabilities for long translation pairs and underestimate those for short phrases. This will cause overfitting and will prevent the system to generalize to unseen data where those short phrases may appear. Filtering away long phrases is also not the best solution, because they may be useful for good translations and cannot be done away with in case they contain non compositional semantic material like idiomatic phrases. Wuebker et al. (2010) used the approach called leaving-one-out (L1O) to deal with overfitting and forced alignment to deal with the errors introduced by incorrect word alignment. The basic idea is to use the trained SMT decoder to re-decode the training data, and then use the decoded result to re-calculate translation

pair probabilities. Since the correct target sentence (i.e. the target side of training data) is not guaranteed to be generated by SMT decoder, forced alignment is used to generate the correct target sentence by discarding all phrase translation candidates which do not match any sequence in the correct target sentence. Since only the correct target sentence can be generated, language model is useless during decoding, so the weight for language model is set to be zero.

Scalable training methods (Perceptron, MIRA and OWL-QN) are used to train the purely discriminative translation model with a large number of features. In order to optimize SMT performance, scalable training tunes the weights to push the best translation candidate upward to be the first one in n-best list. In order to perform scalable training, the n-best candidates should be ranked according to the similarity with the correct target sentence. BLEU is the most natural choice for a similarity measure as it is also the ultimate evaluation criterion. However, BLEU is a document-level metric rather than sentence-level.

Modern phrasal SMT systems (such as Koehn et al. 2003) derive much of their power from being able to memorize and use long phrases. Phrases allow for non-compositional translation, local reordering and contextual lexical choice. However phrases are fully lexicalized, which means they generalize poorly to even slightly out of-domain text. In an open competition (Koehn and Monz 2006) systems trained on parliamentary proceedings were tested on text from 'news commentary' web sites, a very slightly different domain. The nine phrasal systems in the English to Spanish track suffered an absolute drop in BLEU score of between 4.4% and 6.34% (14–27% relative). The treelet system of Menezes et al. (2006) fared somewhat better but still suffered an absolute drop of 3.61%. Clearly there is a need for approaches with greater powers of generalization. There are multiple facets to this issue, including handling of unknown words, new senses of known words etc. At the end of the chapter I will return to the topic of language and translation modeling.

Hybrid and Rule-Based MT Systems

Statistical machine translation (SMT) (Koehn 2010) is currently the leading paradigm in machine translation (MT) research. SMT systems are very attractive because they may be built with little human effort when enough monolingual and bilingual corpora are available. However, bilingual corpora are not always easy to harvest, and they may not even exist for some language pairs. On the contrary, rule-based machine translation systems (RBMT) (Hutchins and Somers 1992) may be built without any parallel corpus; however, they need an explicit representation of linguistic information, whose coding by human experts requires a considerable amount of time.

Rule-Based MT or RBMT for short are by far the mostly used commercial systems still today. This might change in the future. However, it is a fact that SMT has not yet been able to supersede previous work being done on MT which was mainly done in a rule-based fashion.

From a general perspective, hybrid systems are certainly the winning solution. Here I am referring to systems that mix up in a perspicuous manner statistical and non statistical approaches. This can be done in many ways, here are some reported in the literature:

- Using translation memories together with domain trained statistical translation models – this can be done better by using example-based techniques and resources
- Using statistical post-editing before manual supervision with domain trained translation models
- In lack of domain bitexts, providing a dictionary of translation pairs
- Using morphological decomposition for morphologically rich languages (Arabic, German, Italian, French …)
- Using multiword preprocessing in both parallel texts before running language models to reduce semantic uncertainty

When both parallel corpora and linguistic information exist, a hybrid approach may be taken in order to make the most of such resources. In Thurmair (2009) a new hybrid approach is presented which enriches a phrase-based SMT system with resources taken from shallow-transfer RBMT. Shallow-transfer RBMT systems do not perform a complete syntactic analysis of the input sentences, but they rather work with much simpler intermediate representations. Hybridisation between shallow-transfer RBMT and SMT has not yet been explored. Existing hybridisation strategies usually involve more complex RBMT systems and treat them as black boxes, whereas the approach improves SMT by explicitly using the RBMT linguistic resources. They provide an exhaustive evaluation of their hybridisation approach and of the most similar one (Eisele et al. 2008), on the Spanish–English and English–Spanish language pairs by using different training corpus sizes and evaluation corpora.

Rule-based machine translation systems heavily depend on explicit linguistic data such as monolingual dictionaries, bilingual dictionaries, grammars, and structural transfer rules (Hutchins and Somers 1992). Although some automatic acquisition is possible (see Caseli et al. 2006), collecting these data usually requires the intervention of domain experts (mainly, linguists) who master all the encoding and format details of the particular MT system. It could be interesting, however, to open the door to a broader group of non-expert users who could collaboratively enrich MT systems through the web.

Esplà-Gomis et al. (2011) focus on these kinds of dictionaries, which basically have two types of data: paradigms (that group regularities in inflection) and word entries. The paradigm assigned to many common English verbs, for instance, indicates that by adding the ending -ing, the gerund is obtained. Paradigms make easier the management of dictionaries in two ways: by reducing the quantity of information that needs to be stored, and by simplifying revision and validation thanks to the explicit encoding of regularities in the dictionary.

Bilingual dictionaries are the most reused resource from RBMT. They have been added to SMT systems since its early days (Brown et al. 1993). One of the simplest strategies, which has already been put into practice with the Apertium bilingual dictionaries (Tyers 2009; Sanchez-Cartagena and Pérez-Ortiz 2010),

consists of adding the dictionary entries directly to the parallel corpus. In addition to the obvious increase in lexical coverage, Schwenk et al. (2009) state that the quality of the alignments obtained is also improved when the words in the bilingual dictionary appear in other sentences of the parallel corpus. However, it is not guaranteed that, following this strategy, multi-word expressions from the bilingual dictionary that appear in the SL sentences are translated as such because they may be split into smaller units by the phrase-extraction algorithm. Other approaches go beyond simply adding a dictionary to the parallel corpus. For instance, Popovic´ and Ney (2006) propose combining that strategy with the use of hand-crafted rules to reorder the SL sentences to match the structure of the TL.

Although RBMT transfer rules have also been reused in hybrid systems, they have been mostly used implicitly as part of a complete RBMT engine.

For instance, Dugast et al. (2008) show how a PBSMT system can be bootstrapped using only monolingual data and an RBMT engine; RBMT and PBSMT systems can also be combined in a serial fashion (Dugast et al. 2007). Another remarkable study (Eisele et al. 2008) presents a strategy based on the augmentation of the phrase table to include information provided by an RBMT system. In this approach, the sentences to be translated by the hybrid system are first translated with an RBMT system. Then a small phrase table is obtained from the resulting parallel corpus. Phrase pairs are extracted following the usual procedure (Koehn 2010, Sect. 5.2.3) which generates the set of all possible phrase pairs that are consistent with the word alignments. In order to obtain reliable word alignments, they are computed using an alignment model previously built from a large parallel corpus. Finally, the RBMT-generated phrase table is directly added to the original one. Another approach is to generate phrase pairs directly which match either an entry in the bilingual dictionary or a structural transfer rule, thus preventing them from being split into smaller phrase pairs even if they would be consistent with the word alignments. In this way, there is no need for a large parallel corpus from which to learn an alignment model.

España-Bonet et al. (2011) present a system which is guided by a RBMT. In their introduction, they explain why the hybridization is necessary,

> It is well known that rule-based and phrase-based statistical machine translation paradigms (RBMT and SMT, respectively) have complementary strengths and weaknesses. First, RBMT systems tend to produce syntactically better translations and deal with long distance dependencies, agreement and constituent reordering in a better way, since they perform the analysis, transfer and generation steps based on syntactic principles. On the bad side, they usually have problems with lexical selection due to a poor handling of word ambiguity. Also, in cases in which the input sentence has an unexpected syntactic structure, the parser may fail and the quality of the translation decreases dramatically.
>
> On the other side, phrase-based SMT models usually do a better job with lexical selection and general fluency, since they model lexical choice with distributional criteria and explicit probabilistic language models. However, SMT systems usually generate structurally worse translations, since they model translation more locally and have problems with long distance reordering. They also tend to produce very obvious errors, which are annoying for regular users, e.g., lack of gender and number agreement, bad punctuation, etc. Moreover, the SMT systems can experience a severe degradation of performance when applied to corpora different from those used for training (out-of-domain evaluation). (ibid., 554)

The hybrid architecture tries to get the best of both worlds: the RBMT system should perform parsing and rule-based transfer and reordering to produce a good structure for the output, while SMT helps the lexical selection by providing multiple translation suggestions for the pieces of the source language corresponding to the tree constituents. The final decoding accounts also for fluency by using language models, and can be monotonic (and so, fast) because the structure has been already decided by the RBMT component.

System combination, either serial or by a posterior combination of systems' outputs, is a first step towards hybridization. Although it has been shown to improve translation quality, the combination does not represent a real hybridization since systems do not interact among them (see Thurmair 2009) for a classification of HMT architectures. In the case of actual interdependences, one of the systems in action leads the translation process and the other ones strengthen it. Much work has been done in building systems in which the statistical component is in charge of the translation and the companion system provides complementary information. For instance, Eisele et al. (2008) and Chen and Eisele (2010) introduce lexical information coming from a rule-based translator into an SMT system, in the form of new phrase pairs for the translation table. In both cases results are positive on out-of-domain tests.

The opposite direction is less explored. In this case, the RBMT system leads the translation and the SMT system provides complementary information. Habash et al. (2009) enrich the dictionary of a RBMT system with phrases from an SMT system (see also Alkuhlani and Habash 2011). Federmann et al. (2010) use the translations obtained with a RBMT system and substitute selected noun phrases by their SMT counterparts. Globally, their results improve the individual systems when the hybrid system is applied to translate into languages with a richer morphology than the source. In Figure 6.2 below there is a pipeline for a generic Hybrid System that combines a Rule Based approach with Statistical Models.

Specific Issues in Hybrid MT

A number of specific issues are dealt with inside this framework, even though they may certainly be regarded as general problems of SMT. In particular, the treatment of English particle and of function words, is a topic that has developed into a number of interesting techniques.

Morpheme-based SMT system (SMTm) a second variant of the SMT system was used to address the rich morphology of Basque. In this system, words are split into several morphemes by using a Basque morphological analyzer/lemmatizer. The aim is to reduce the sparseness produced by the agglutinative nature of Basque and the small amount of parallel corpora. Adapting the baseline system to work at the morpheme level mainly consists of training Moses on the segmented text. The SMT system trained on segmented words will generate a sequence of morphemes. So, in order to obtain the final Basque text from the segmented output, a word-generation post-process is applied. Details on this system can be found in (Labaka 2010).

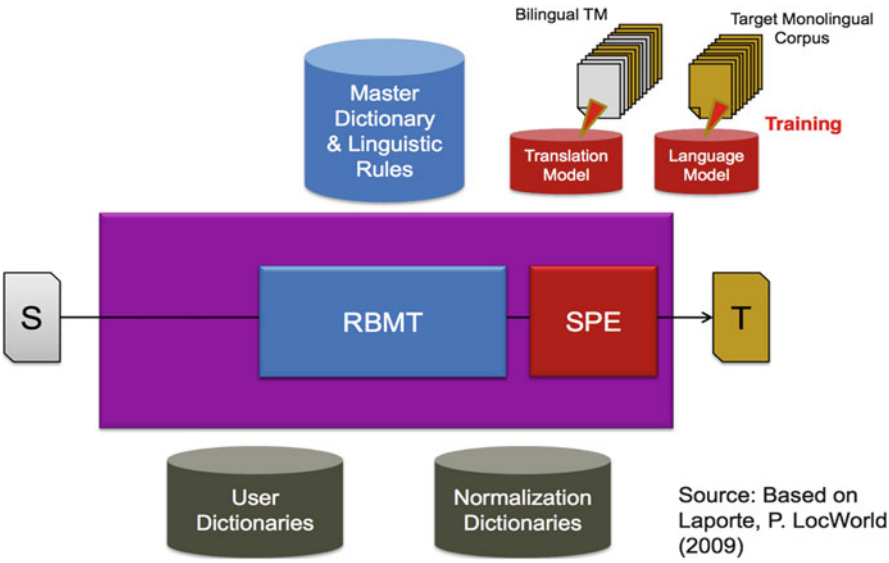

Fig. 6.2 Hybrid Systems

Matxin is another interesting hybrid rule-based system, an open-source Spanish-Basque RBMT engine (Alegria et al. 2007), following the traditional transfer model. The system consists of three main components: (1) analysis of the source sentence into a dependency tree structure; (2) transfer from the source language dependency tree to a target language dependency structure; and (3) generation of the output translation from the target dependency structure. The SMatxinT architecture is based on the three following principles: (1) generally, the final translation should be based on RBMT system's syntactic rearrangements; (2) the hybrid system must have the chance of using SMT-based local translations to improve lexical selection; and (3) it should be able to recover from potential problems encountered in the analysis, using longer SMT translations.

Phillips 2011 discusses other ways of overcoming shortages of SMT, introducing ways to incorporate the context for each translation instance. To overcome these deficiencies Phillips proposes to model each instance of translation. An instance of translation is the realization of a source and corresponding target phrase at one specific location in the corpus. He defines his method as follows:

> An instance of translation is the realization of a source and corresponding target phrase at one specific location in the corpus. We score each translation instance with a series of features that examine the alignment, context, genre, and other surroundings. Our model then combines these translation instances in a weighted summation. This approach conveniently side-steps the challenges of estimation sparsity because our model is not based on relative frequency estimates. The weighting of translation instances relates to methods for domain-adaptation of SMT models, but our implementation is fundamentally different in that we do not alter or re-weight the training data. Instead, our model directly embodies the notion that

> not all translations are equal and individually evaluates the relevance of each translation instance… Evidence for a translation unit θ will generally be present at multiple locations within the training data. The features for θ operate over this set of translation instances and are generally computed as relative frequencies. A common feature, for example, is the number of times source and target instances are aligned divided by the total occurrences of source instances.
>
> Our model for translation is fundamentally different in that our translation units are not abstract phrase pairs or grammar rules … the core component of our model is a feature function which allows the user to easily add new sources of knowledge to the system. However, our feature function ϕ evaluates one specific instance of translation instead of scoring the entire set of translation instances.

In fact, they produced this new statistical model because they wanted explicitly to incorporate ideas coming from EBMT,

> For illustration, consider that the translation instances for a given phrase pair occur in a variety of sentences within the training data. Some instances may include an inconsistent word alignment from within the selected phrase pair to a word in the remainder of the sentence. Our model allows us to learn from these translation instances, but discount them by including a feature in ϕ which measures the likelihood of the phrasal alignment given the words outside the phrase pair. This differs from the standard SMT approach where phrase alignment is a binary decision. The same principle also applies if we want to include additional non-local information such as genre or context within the model. A traditional SMT model requires new translation units conditioned on the extra information whereas our approach incorporates the extra information as features of ϕ and calculates a score over all instances.
>
> One of the motivations for this model was to combine ideas from Statistical MT and Example-Based MT. Many EBMT systems rely on heuristics and lack a well-defined model, but our per-instance modeling is generally reflective of an 'EBMT approach.'

English Particles and Function Words

Ma Jianjun et al. (2011) address the problem of correctly translating English particles (adverbs and prepositions) into Chinese. They introduce POS tags in the corpus, and thus tags become an important feature for the Maximum Entropy model. For that purpose, they use the Stanford tagger. However, in order to improve the results, they have to proceed to some post-processing operation with rules that take into account typical phrasal verb collocations from a manually built collocation bank.

In practice, many function words do not have the exact counterparts in the other language and will not align to any words (i.e. align to NULL) in the results of word alignment. Furthermore, due to the high frequencies of function words, they could be associated with any content words to form bilingual phrases which might be quite noisy.

Consequently, many target function words may be either missing or inappropriately generated in the translations. This not only degrades the readability but also impairs the quality of content word translations.

The incompleteness of target function word generation seems to be mainly caused by the noisy translation knowledge automatically learnt based on word alignment.

In particular, some words serve to express (language-specific) grammatical relations only, and thus they may have no counterpart in another language. This problem is nicely treated in Liu et al. (2011). They divide up words into two subcategories: spurious and non spurious words. The first type should be aligned to a null token. For example the Chinese words "hen" and "bi" have no counterparts on the other side, neither do the English words "does" and "to". To deal with the spurious words in sentence pairs, IBM models 3, 4 and 5 (Brown et al. 1993) introduce a special token null, which can align to a source/target word. Fraser and Marcu (2007a) proposed a new generative model called LEAF, which is similar to the IBM models, in which words are classified into three types instead of two: spurious words, head words (which are the key words of a sentence) and non-head words (modifiers of head words).

"Spurious" words usually have no counterpart in other languages, and are therefore a headache in machine translation. The authors propose a novel framework, called skeleton-enhanced translation, in which a conventional SMT decoder can boost itself by considering the skeleton of the source input and the translation of such skeleton. The skeleton of a sentence is the sentence with its spurious words removed. Two models for identifying spurious words, are introduced. The first one is a context-insensitive model, which removes all tokens of certain words. The second one is a context-sensitive model, which makes separate decision for each word token. The authors also elaborate two methods to improve a translation decoder using skeleton translation. One is skeleton-enhanced re-ranking, which re-ranks the n-best output of a conventional SMT decoder with respect to a translated skeleton. Another is skeleton-enhanced decoding, which re-ranks the translation hypotheses of not only the entire sentence but any span of the sentence. Their experiments show significant improvement (1.6 BLEU) over the state-of-the-art SMT performance.

These two methods are generative models for word alignment. Nevertheless they cannot be used directly in the conventional log-linear model of statistic machine translation (SMT). The conventional phrase-based SMT captures spurious words within the phrase pairs in the translation table. As Liu et al. (2011) note, the existence of spurious words in training data leads to a certain kind of data sparseness. For example, "na bi qian" and "na xie qian" share the same translation ("that money"). If the spurious words ("bi" and "xie") are removed, then the two entries in translation table, and the associated statistics, can be combined into one. However, while spurious words lead to the harmful effect of data sparseness, they are useful in certain aspects in translation. To cope with this problem, as automatic word alignment is far from perfect, in keeping a high precision of spurious word deletion. It is stipulated that a word token is not to be removed unless the model assigns a high probability to the deletion decision.

Correct translation of function words into Chinese is discussed in a paper by Cui et al. (2011). They have been interested in the subject because "… function words play an important role in sentence structures and express grammatical relationships with other words"(ibid., 139). Most statistical machine translation (SMT) systems do not pay enough attention to translations of function words which are noisy due to data sparseness and word alignment errors. Their method is designed to separate

the generation of target function words from target content words in SMT decoding. With this method, the target function words are deleted before the translation modeling while in SMT decoding they are inserted back into the translations. To guide the target function words insertion, a new statistical model is proposed and integrated into the log-linear model for SMT. This can lead to better reordering and partial hypotheses ranking. As shown by experimental results, their approach improves the SMT performance significantly on Chinese-English translation task.

For example, when considering the top eight function words, about 63.9% of Chinese function word occurrences are not aligned to any English words and about 74.5% of Chinese sentences contain at least one unaligned Chinese function word. On the English side, about 36.5% of English function word occurrences are not aligned to any Chinese words and about 88.8% of English sentences contain at least one unaligned English function word.

Combining Translation Memories with EBMT

Even though over the past two decades, machine translation has shown very promising results, a large number of languages exist which suffer from the scarcity of parallel corpora, e.g. Indic languages, sign languages etc. SMT approaches have yielded low translation quality for these poorly resourced languages (Khalilov et al. 2010). It is often the case that domain-specific translation is required to tackle the issue of scarce resources, but it can still suffer from very low accuracy within the SMT framework, even for homogeneous domains (Dandapat et al. 2010). Although SMT and EBMT are both data-driven approaches to MT, both of them have their own advantages and limitations. Typically, an SMT system works well with significant amounts of training data. In contrast, an EBMT approach can be developed with a limited example-base (Somers 2003); also, as with any other data-driven system, an EBMT system works well when training and test sets are quite close in nature. This is because EBMT systems reuse the segments of test sentences that can be found in the source side of the example-base at runtime (see Brown 1996). Keeping these points in mind is important in order to develop an MT system of reasonably good quality based on limited amounts of data. In this direction, they examine different EBMT approaches which can handle the problem of data sparseness. It is often the case that EBMT systems produce a good translation where SMT fails and vice versa. In order to harness the advantages of both approaches, they use a careful combination of both EBMT and SMT to improve translation accuracy.

Two alternative approaches are adopted to tackle the above problems. First there is a compiled approach to EBMT which essentially produces translation templates during the training stage, based on the description in (Cicekli and Güveniri 2001). The second attempt presents a novel way of integrating translation memory (TM) into an EBMT system. Starting with the user's TM as a training set, additional sub-sentential translation units (TUs) are extracted based on the word alignments produced by an SMT system. These sub-sentential TUs are used both for alignment and recombination after the closest matching example to the input is found in the matching stage of our EBMT system.

Simard et al. (2007) note that TM has some notable advantages over most data-driven MT systems. The most obvious is its ability to translate predictably and (near-) perfectly any input that it has seen previously. Another quality of TM is its ability to find approximate matches and to let the user adapt system behavior to his/her own tolerance to errors by fixing the similarity threshold on such matches; in other words, TM's benefit from a highly effective confidence estimation mechanism. If machine translation is to be integrated successfully in the CAT environment, it should begin by catching up with TM on these aspects. This requires two things: (1) the MT system should behave more like a TM in the presence of high-similarity matches. In practice, this can be achieved by combining the two technologies, i.e. by building a combination MT system that incorporates a TM component. And (2) just like existing TM systems, the combined MT system should provide the user with means to filter out translations that are less likely to be useful. It has sometimes been proposed (see e.g. Heyn 1996) that MT should be used within a CAT environment only when the TM fails to retrieve something useful. Unfortunately, this has the effect of relegating the MT system to the task of translating only the sentences that are most unlike previously seen ones. For data-driven systems, this means translating only the "harder" sentences and missing the chance to do a better job than the TM. The reason why MT is often treated as a last resort lies in the fact that translators tend to see its performance as unpredictable and, as a result, overly likely to waste their time.

Example-based MT has the problem of coverage and the fragments used are not decomposable so they have limited flexibility. This may be improved by Translation Memories made available by users in a specific domain, such as the just released databank of TMs from JRC (see above).

In 2003, the idea of combining knowledge coming from Translation Memories, which is very domain localised, and EBMT was not yet clearly formulated. Even the organizers, Carl and Way, of the corresponding workshop admit this. While translation memory systems are used in restricted domains, SBMT systems require training on huge, good quality bilingual corpora. As a consequence TMs can hardly be applied as a general purpose solution to MT, and SBMT as yet cannot produce complex translations to the desired quality, even if such translations are given to the system in the training phase. EBMT seeks to exploit and integrate a number of knowledge resources, such as linguistics and statistics, and symbolic and numerical techniques, for integration into one framework. In this way, rule-based morphological, syntactic and/or semantic information is combined with knowledge extracted from bilingual texts which is then re-used in the translation process.

However, it is unclear how one might combine the different knowledge resources and techniques in an optimal way. In EBMT, therefore, the question is asked: what can be learned from a bilingual corpus and what needs to be provided manually? Furthermore, it is uncertain how far the EBMT methodology can be pushed with respect to translation quality and/or translation purpose. Finally, one wonders what the implications and consequences are for the size and quality of the reference translations, (computational) complexity of the system, sizeability and transportability, if such an approach is taken.

Sanchez-Cartagena et al. (2011), extensively evaluate a new hybridisation approach. It consists of enriching the phrase table of a phrase-based statistical

machine translation system with bilingual phrase pairs matching transfer rules and dictionary entries from a shallow-transfer rule-based machine translation system. The experiments conducted show an improvement in translation quality, specially when the parallel corpus available for training is small (see also Sanchez-Martinez and Forcada 2009), or when translating out-of-domain texts that are well covered by the shallow-transfer rule-based machine translation system.

In their paper Dandapat et al. (2011) address the issue of applying example-based machine translation (EBMT) methods to overcome some of the difficulties encountered with statistical machine translation (SMT) techniques. They adopt two different EBMT approaches and present an approach to augment output quality by strategically combining both EBMT approaches with the SMT system to handle issues arising from the use of SMT. They use these approaches for English to Turkish translation using the IWSLT09 dataset. Improved evaluation scores (4% relative BLEU improvement) were achieved when EBMT was used to translate sentences for which SMT failed to produce an adequate translation.

Ebling et al. (2011) present Marclator and its ability to chunk based on the so-called "Marker Hypothesis" (Green 1979), which is a psycholinguistic hypothesis stating that every language has a closed set of elements that are used to mark certain syntactic constructions. Marclator system segments both the training and the test data into chunks, where the set of elements includes function words and bound morphemes (-ing as an indicator of English progressive-tense verbs). The interesting point is that Marclator chunking module solely considers function words as indicators of chunk boundaries, and head words are included in their inflected forms. In fact, each function word (Marker word) triggers the opening of a new chunk, provided that the preceding chunk contains at least one non-Marker word.

Chunk example: He was | on the bus

Typical problems inherent in this approach are the chunks of an input sentence that often cannot be found in the example base. So the goal is to increase the chunk coverage of a system. Gough and Way (2003) extended a precursor to Marclator by including an additional layer of abstraction: producing generalized chunks by replacing the Marker word at the beginning of a chunk with the name of its category.

For example: of a marathon | <PREP> a marathon

OPENMATREX is another system based on the marker hypothesis reported lately in Banerjee et al. 2011. As the authors comment in the conclusion, "OpenMaTrEx comprises a marker-driven chunker, a collection of chunk aligners, tools to merge (hybridise") marker-based and statistical translation tables, two engines a simple proof-of-concept monotone "example-based" recombination engine and a statistical decoder based on Moses, and support for automatic evaluation. It also contains support for "word packing" to improve alignment."(ibid. 14) The performance of the system shows improvements over the purely statistical mode.

In CMU-EBMT II, the system generalizes both the training and the test set: it recursively replaces words and phrases that are part of an equivalence class with the corresponding class tags. Syntactic classes are applied before semantic classes. In training data, a generalization is performed only if a member of a particular equivalence class is found in both the SL and the corresponding TL sentence. Eventually, in CMU-EBMT III the test set has all members of an equivalence class that are replaced recursively.

The matching process is equivalent to that of the purely lexical CMU-EBMT system, with the apparent difference that here, two matching levels: a lexical and a generalized one exist. The alignment proceeds in the same way as in CMU-EBMT. Following this, the rules that were stored during the generalization of the input sentence are applied in reverse so as to transform the generalized TL fragments into word form TL fragments. The system carries out translations by matching chunks first. If the system does not find a chunk in the example base, it proceeds to replace the Marker word at the beginning of a chunk with its corresponding Marker tag and to search for the resulting generalized chunk in the example base (if this attempt fails, the system reverts to word-by-word translation).

One major source of errors is chunk-internal boundary friction. Boundary friction is normally caused by combining two separate translation units that do not agree in grammatical case with the introduction of Marker-based templates. It can also take place within a single chunk, i.e., when a Marker word is inserted that does not agree with the grammatical properties of the rest of the chunk. In the case of translating from English to German, inserting TL Marker words in a context-insensitive manner (as is done in System 1) is error prone. Due to the morphological richness of German, an English Marker word can correspond to multiple word forms of the same lemma on the German side e.g., English Marker word "are" can correspond to German Marker words "bist, sind" and "seid". Example: for "are you sure … /sind du sicher …" where the chunk-internal boundary friction causes a combination of "sind" and "du" which is grammatically incorrect.

Eventually I want to include a short note on Graph-Based Learning approaches, which will be explained in a section below – that can be likened to a probabilistic implementation of translation memories (Maruyana and Watanabe 1992; Veale and Way 1997). Translation memories are (usually commercial) databases of segment translations extracted from a large database of translation examples. They are typically used by human translators to retrieve translation candidates for subsequences of a new input text. Matches can be exact or fuzzy; the latter is similar to the identification of graph neighborhoods in our approach. However, the GBL scheme propagates similarity scores not just from known to unknown sentences but also indirectly, via connections through other unknown sentences. Marcu 2001 reported the combination of a translation memory and statistical translation; however, this is a combination of word-based and phrase-based translation predating the current phrase-based approach to SMT.

Automatic Post-editing for Translator CAT Tools

The translation quality of MT has been improving but has not reached an adequate level compared with human translation. As such, manual evaluation and post-editing constitute an essential part of the translation processes. To make the best use of MT, human translators are urged to perform post-editing efficiently and effectively. Therefore there is a huge demand for MT to alleviate the burden of manual post-editing.

Alleviating the burden for human post-editing is the aim of research efforts in the direction of producing automatic procedures that work possibly on the basis of the output of STM or RBMT and TM or at least strongly domain limited bitexts. In fact "bitext"

is not synonymous with parallel corpora, as Tiedemann 2011 notes. Suzuki (2011), working for Toshiba, has built a quality prediction model with regression analysis for Japanese English and viceversa APE, where confidence estimation (CE) is considered as also (Specia et al. 2009a; 2009b) do, by estimating a continuous translational quality score for each sentence, using PLS (Partial Least Squares) regression analysis. Since Rule-based MT (RBMT) is generally more stable in translation quality than SMT, it can make it easier to integrate the post-editing into the translation processes. This, however, is also a weak point of RBMT because post-editors are forced to repeatedly correct the same kind of errors made by MT systems (see Roturier 2009). Statistical post-editing (SPE) techniques have been successfully applied to the output of Rule Based MT (RBMT) systems. In the computing assisted translation process with machine translation (MT), post-editing costs time and efforts on the part of human. To solve this problem, some have attempted to automate post editing. Post-editing isn't always necessary, however, when MT outputs are of adequate quality for human. This means that we need to be able to estimate the translation quality of each translated sentence to determine whether post-editing should be performed. While conventional automatic metrics such as BLEU, NIST and METEOR, require the golden standards (references), for wider applications we need to establish methods that can estimate the quality of translations without references. The paper presents a sentence-level automatic quality evaluator, composed of an SMT phrase-based automatic post-editing (APE) module and a confidence estimator characterized by PLS regression analysis. It is known that this model is a better model for predicting output variable than a normal multiple regression analysis when the multicollinearity exists between the input variables. Experiments with Japanese to English patent translations show the validity of the proposed methods.

Recognizing that SMT is better suited to correct frequent errors to appropriate expressions, some (Simard et al. 2007; Lagarda et al. 2009) have proposed to use SMT for an automatic post-editor and built an automatic post-editing module, where MT outputs are regarded as source sentences and manually post-edited/translated results as target sentences.

Béchara et al. (2011) investigate the impact of SPE on a standard Phrase-Based Statistical Machine Translation (PB-SMT) system, using PB-SMT both for the first-stage MT and the second stage SPE system. Their results show that, while a naive approach to using SPE in a PB-SMT pipeline produces no or only modest improvements, a novel combination of source context modeling and thresholding can produce statistically significant improvements of 2 BLEU points over baseline using technical translation data for French to English.

Simard et al. (2007a) train a "mono-lingual" PB-SMT system (the Portage system) on the output of an RBMT system for the source side of the training set of the PB-SMT system and the corresponding human translated reference. A complete translation pipeline consists of a rule-based first-stage system, whose output on some (unseen) test set, in turn, is translated by the second-stage "mono-lingual" SPE system. Simard et al. (2007a) present experiments using Human Resources and Social Development (HRSDC) Job Bank1 French and English parallel data. They found that in combination, the RBMT system post-edited by the PB-SMT system performed

significantly better than each of the individual systems on their own. Simard et al. (2007a) also tested the SPE technique with Portage PB-SMT both as first-stage MT and as second stage SPE system (i.e. Portage post-editing its own output) and reported that nothing could be gained. In a number of follow-up experiments, Simard et al. (2007b) used an SPE system to adapt RBMT-systems to a specific domain, once again using Portage in the SPE phase. Adding the SPE system produced BLEU score increases of about 20 points over the original RBMT baseline.

SPE was also applied in an attempt to improve Japanese to English patent translations. Teramusa (2007) uses RBMT to translate patent texts, which tend to be difficult to translate without syntactic analysis. Combining RBMT with SPE in the post-editing phase produced an improved score on the NIST evaluation compared to that of the RBMT system alone. Dugast et al. (2007) report research on combining SYSTRAN with PB-SMT systems Moses and Portage. Comparison between raw SYSTRAN output and SYSTRAN + SPE output shows significant improvements in terms of lexical choice, but almost no improvement in word order or grammaticality. Dugast et al. (2009) trained a similar post-editing system with some additional treatment to prevent the loss of entities such as dates and numbers.

Oflazer and El-Kahlout (2007) explore selective segmentation-based models for English to Turkish translation. As part of their experiments they present a short section at the end of the paper on statistical post-editing of an SMT system, which they call model iteration. They train a post-editing SMT model on the training set decoded by the first stage SMT model and iterate the approach, post-editing the output of the post-editing system. BLEU results show positive improvements, with a cumulative 0.46 increase after two model iterations. It is not clear whether the result is statistically significant. The experiments follow the statistical post-editing design of Simard et al. (2007a), where the output of a first-stage system is used to train a mono-lingual second stage system, that has the potential to correct or otherwise improve on (i.e. post-edit) the output of the first-stage system. The experiments use PB-SMT systems throughout both stages. The objective is to investigate in more detail whether and to what extent state-of-the-art PBSMT technology can be used to post-edit itself, i.e. its own output.

Blain et al. (2011) report on work on post-editing by Systran and Symantec where they define what they call a Post-Editing Action (PEA) typology on the basis of a detailed analysis of errors, which we report here below (166–167):

Noun-Phrase (NP) – related to lexical changes.

- Determiner choice – change in determiner
- Noun meaning choice – a noun, replaces another noun, changing its meaning
- Noun stylistic change – a synonym replaces a noun (no meaning change)
- Noun number change
- Case change
- Adjective choice – change in adjective choice for better fit with modified noun
- Multi-word change – multiword expression change (meaning change)
- NP structure change – structure change of NP but the sense is preserved

- Verbal-Phrase (VP) – related to grammatical changes
- Verb agreement – correction of agreement in verb
- Verb phrase structure change
- Verb meaning choice – a verb replaces another verb, changing its meaning
- Verb stylistic change – a synonym replaces a verb.

Preposition change

- Co-reference change – generally through introduction/removal of a pronoun, or change of a definite to possessive determiner
- Reordering – repositioning of a constituent at a better location (adjective, adverb)
- PE Error – Post-editor made a mistake in his review
- Misc style – unnecessary stylistic change
- Misc – all PEAs that we cannot classify

The UNL: Universal Networking Language

In a paper online, Alansary et al. present the UNL concisely and report some recent data. One of the challenging missions that the UNL system has to face is to translate the Encyclopedia of Life Support System (EOLSS) which is the largest on-line Encyclopaedia; it includes more than 120,000 web pages and it increases constantly. The translation results are reported as reaching a morphological accuracy of 90%, a syntactic accuracy of 75% and a semantic accuracy of 85%. The adopted approach in the translation in this abstract follows a different way, it translates from a semantically-based Interlingua to different human languages. The UNL (see Adly and Alansary 2009) has been introduced by the United Nations University, Tokyo, to facilitate the transfer and exchange of information over the internet. The semantic representation is an artificial language which describes the meaning of sentences in terms of the schema of semantic nets. It aims to represent all sentences that have the same meaning in all natural languages using a single semantic graph. Once this graph is built, it is possible to decode it to any other language. UNL is used not only in machine translation and other natural language processing tasks, but also in a wide variety of applications ranging from e-learning platforms to management of multilingual document bases. Working at the semantic level, the UNL is language-independent: in particular, it follows the schema of semantic nets-like structure in which nodes are word concepts and arcs are semantic relations between these concepts. In this scheme, a source language sentence is converted to the UNL form using a tool called the EnConverter. EnConverter is a language independent parser that provides synchronously a framework for morphological, syntactic and semantic analysis. Subsequently, the UNL representation is converted to the target language sentence by a tool called the DeConverter. The DeConverter is a language independent generator that provides a framework for syntactic and morphological generation as well as co occurrence-based word selection for linguistic collocations.

It can deconvert UNL expressions into a variety of native languages, using a number of linguistic data such as Word Dictionary, Grammatical Rules and Co-occurrence Dictionary of each language. UNL's main task and purpose is translating The Encyclopedia Of Life Support Systems (EOLSS), because it provides a useful body of knowledge which should reach all peoples in their languages and in a way that fits their cultural backgrounds. UNL can do both: reproduce EOLSS knowledge in peoples' native languages, and enable them to explore it according to their cultural backgrounds. The UNL task is to make the entire EOLSS available in multiple languages starting with the six official languages of UNESCO. This task involves a two-step process: the first step is enconverting (encoding) the content of EOLSS from English into UNL (UNLization process); and the second is deconverting (decoding) EOLSS content from UNL into natural languages.

Combining Syntax, EBMT and a Transfer Approach to MT

Vandeghinste and Martens (2010) present another interesting option, in a number of papers in which the authors describe a system PaCo-MT that uses a transfer approach where syntax and examples are combined in a stochastic model. We quote from the Vandeghinste and Martens (2010) paper describing, "… the transfer component of a syntax-based Example-based Machine Translation system. The source sentence parse tree is matched in a bottom-up fashion with the source language side of a parallel example treebank, which results in a target forest … sent to the target language generation component." The novelty of the approach described in this paper was the bottom-up policy as opposed to the top-down one, in the choice of source sentence parse tree. Translations are example-based, in that "… as it uses a large set of translation examples (a parallel corpus) as training data to base its decisions on and it is *syntax-based* as the data in the parallel corpus is annotated with syntactic parse trees, both on the source and the target side. Input sentences are syntactically analysed, and the system generates target language parse trees where all ordering information is removed." The system uses a parser for the source language to parse the source side of the parallel corpus as well as the input sentence to feed the translation engine. The target language parser is only used to preprocess the target parallel corpus. The parallel treebank has also been commented upon in Tiedemann and Kotzé (2009), it is word aligned using GIZA++, and node aligned using a discriminative approach to tree alignment (Tiedemann 2011).

As the authors comment, using a syntax-based translation unit is like using a rule-based approach. In fact, the PaCo-MT system combines a stochastic example-based transfer system with the data-driven tree-to-tree based approach, transducing the source parse tree into a set of target language parse trees. This is done without node ordering, and reordering is done by a discriminative model for tree alignment. In this way, rule-based strengths are combined with PBSMT systems: in particular, the target tree-based language model is generated using a probabilistic context-free grammar

based on large monolingual treebanks which rather than reordering words or phrases, it addresses parse trees. We address more of these problems in the next section.

Syntax Based Approaches: From Hierarchical to SBSMT

Accurate translation may ensue from SMT and EBMT but there is no way to control the performance of such systems to obtain a 100% accurate translation all the time. So improvements may only come from external knowledge made available to the system either at runtime, producing some preprocessing and new models, or at the end of the computation, producing some postprocessing. For sometime the introduction of syntactic information in the training process did not seem to produce any improvement in the performance of STM. However, a number of papers appearing lately show that this is not always the case. In particular in language pairs which require heavy reordering, and/or have totally different grammatical structures, syntactic information seems particularly useful. The need for reordering in some language pairs is paramount and cannot be limited to local phrases. Syntax may provide means for an accurate reordering step. Syntax may also check for appropriate insertion of function words and their wordforms – in case of amalgamated function words like articulated prepositions in German and Romance languages.

- Language may have the problem of pro-dropping subject and object (like Japanese) or just subject as most Romance languages do;
- The most typical problem is semantic and word sense ambiguity that requires disambiguation: this may be done only by restricting the language model to a specific translation domain where the appropriate sense is usually easily capture. Or else a full-fledged words-sense disambiguation algorithm must be in place;
- Languages may use tenses differently or have more/less tenses – like perfect in English, and "imperfetto" in some Romance languages, simple past and "passato prossimo" versus "passato remoto" in Italian;
- Idioms may be difficult to trace in complete phrases (see Wehrli 2007, Wehrli et al. 2009) on the subject);
- Many transformations can be best explained in syntactic terms – see examples below;
- Syntactic annotation on the source input adds additional knowledge
- Syntactic annotation on the target output aids grammatical output

Here are some attempts at using syntax-based models:

- String to tree based translation systems (Yamada and Knight 2001; Galley et al. 2006; Marcu et al. 2006; Shen et al. 2008; Chiang et al. 2009)
- Using syntactic chunks (Schafer and Yarowsky 2003)
- Using syntactic features (Koehn and Knight 2003; Och and Ney 2003)

– Tree-to-string based translation systems (Quirk et al. 2005; Liu et al. 2006; Huang et al. 2006; Mi et al. 2008)
– Tree-to-tree based translation systems (Eisner 2003; Ding and Palmer 2005; Cowan et al. 2006; Zhang et al. 2007; Liu et al. 2009)

Early SMT syntactic models had worse results than PBSMT because phrase pairs limited to corresponding complete syntactic units were harmful for translation. Some of the advantages of SBMT are:

– Better overall handling of word order
– Better at translating discontinuous phrases (E.g. as X as Y- aussi X que Y)
– Especially advantageous for handling typologically different languages
– Fast and steady improvement in recent years

Syntax-based approaches for Machine Translation (MT) have gained popularity in recent times because of their ability to handle long distance reorderings (Wu 1997; Yamada and Knight 2002; Quirk et al. 2005; Chiang 2005), especially for divergent language pairs such as English-Hindi (or English-Urdu). Languages such as Hindi are also known for their rich morphology and long distance agreement of features of syntactically related units. Employing techniques that factor the lexical items into morphological factors can handle the morphological richness. The same applies to Arabic (see El Kholy and Habash 2010).

The first problem that SBMT aimed to solve was the issue of reordering, i.e. learning how to transform the sentence structure of one language into the sentence structure of another, in a way that is not tied to a specific domain or sub-domains, or indeed, sequences of individual words. An early attempt at greater generality in a purely phrasal setting was the alignment template approach (Och and Ney 2004). Newer approaches include formally syntactic (Chiang 2005), and linguistically syntactic approaches (Quirk et al. 2005; Huang et al. 2006; Wang et al. 2010).

The other fundamental issue SBMT targets, is extraposition and long distance movement which still pose a serious challenge to syntax-based machine translation systems. Even if the search algorithms could accommodate such syntactic discontinuities, we need appropriate models to account for such phenomena. Also if the system extracts extraposition templates, they may prove too sparse and brittle to accomodate the range of phenomena.

String models are popular in statistical machine translation. Approaches include word substitution systems (Brown et al. 1993), phrase substitution systems (Koehn et al. 2003; Och and Ney 2004), and synchronous context-free grammar systems (Wu and Wong 1998; Chiang 2005; Wong et al. 2005; Huang et al. 2009), all of which train on string pairs and seek to establish connections between source and target strings. By contrast, explicit syntax approaches seek to model directly the relations learned from parsed data, including models between source trees and target trees (Gildea 2003; Eisner 2003; Melamed 2004; Cowan et al. 2006), source trees and target strings (Quirk et al. 2005; Huang et al. 2006), or source strings and target trees (Yamada and Knight 2001; Galley et al. 2004). A strength of phrase models is that they can acquire all phrase pairs consistent with computed word alignments (Lopez

and Resnik 2006), concatenate those phrases together, and re-order them under several cost models. An advantage of syntax-based models is that outputs tend to be syntactically well-formed, with re-ordering influenced by syntactic context and function words introduced to serve specific syntactic purposes.

Generally speaking, syntactic models outperform string models for the simple reason that their output is still syntactically acceptable even for bad translations that may be semantically wrong, whereas the former produces bad translations that are also grammatically totally incorrect.

Hierarchical MT

In 2005 the first SMT system that uses hierarchical phrase-based decoding (HPBSMT) is presented (Chiang 2005), and is shown to improve the performance of phrase-based systems at least for all those concerned with Chinese. HPBSMT extends the PBSMT by allowing the use of non-contiguous phrase pairs. It incorporates reordering rules and in some way also the recursive structure of the sentence, implicitly adopting in this way a linguistic approach without including any linguistic representation of the data. To make the model sensitive to the syntax structure, a constituent feature was integrated into the translation model with the soft constraint method. It was defined as follows: it gains 1 for rules whose source side respect syntactic phrase boundary in the parse tree, and 0 otherwise. However, it did not achieve statistically significant improvement in the experiment. Marton and Resnik (2008) (hence M&R 2008) thought that different syntactic types may play different roles in the translation model. However, (Chiang 2005)'s method did not treat them discriminatively. They then defined soft constraint features for each constituent type based on the observation of this phenomenon. Their experiments showed that some constituent features significantly improved the performance, but others didn't. It is an interesting question whether all these constituent type models can work together efficiently. Although M&R 2008 did not give the experiments to support the positive answer. Chiang (2005) had already provided the evidence that their constituent models could not work together. (Chiang et al. 2008) thought one of its reasons were the limitations of MERT (Och 2003) with many features. We explore the topic of soft constraints more below.

HPBSMT is usually described as being formally similar to a syntactic model without linguistic commitments, in contrast with syntactic decoding which uses rules with linguistically motivated labels. However, as remarked in Hoang and Koehn (2010) – hence HK2010, the decoding mechanism for both hierarchical and syntactic systems are identical and the rule extraction are similar. Hierarchical and syntax statistical machine translation have made great progress in the last few years and now represent the state of the art in the field. Both use synchronous context free grammar (SCFG) formalism, consisting of rewrite rules which simultaneously parse the input sentence and generate the output sentence. The most common algorithm for decoding with SCFG is currently CKY+ with cube pruning, which works for

both hierarchical and syntactic systems, as implemented in Hiero (Chiang 2005), Joshua (Li et al. 2009), and Moses (Hoang et al. 2009). Again as commented by HK2010, simple HPBSMT have the advantage of ensuring broad coverage to their representations, but run the risk of using a rule for an inappropriate situation.

Most existing alignment methods simply consider a sentence as a sequence of words (Brown et al. 1993), and generate phrase correspondences using heuristic rules (Koehn et al. 2003). Some studies incorporate structural information into the alignment process *after* this simple word alignment (Quirk et al. 2005; Cowan et al. 2006). However, this is not sufficient because the basic word alignment itself is not good.

On the other hand, syntactic models have been proposed which use structural information from the beginning of the alignment process. Watanabe et al. (2000) and Menezes and Richardson (2001) proposed a structural alignment method. These methods use heuristic rules when resolving correspondence ambiguities. Yamada and Knight (2001) and Gildea (2003) proposed a tree-based probabilistic alignment methods. These methods reorder, insert or delete sub-trees on one side to reproduce the other side, but the constraints of using syntactic information is often too rigid. Yamada and Knight flattened the trees by collapsing nodes. Gildea cloned sub-trees to deal with the problem.

Rewrite rules in hierarchical systems have general applicability as their non-terminals are undecorated, giving hierarchical system broad coverage. However, rules may be used in inappropriate situations without the labeled constraints. The general applicability of undecorated rules create spurious ambiguity which decreases translation performance by causing the decoder to spend more time sifting through duplicate hypotheses. Syntactic systems make use of linguistically motivated information to bias the search space at the expense of limiting model coverage. The main problem to solve when using syntactic representation is the poor coverage of syntactically encoded translation rules and as a result the decoding phase has a low number of translation pairs.

Eventually, the ability to incorporate both source and target syntactic information in tree-to-tree models are believed to have a lot of potential to achieve promising translation quality. However, they are affected by rigid syntactic constraints and this may be the reason that conventional tree-to-tree based translation systems haven't shown superiority in empirical evaluations. We address more on this topic below.

Syntactic labels from parse trees can be used to annotate non-terminals in the translation model. This reduces incorrect rule application by restricting rule extraction and application. However, as noted in (Ambati and Lavie 2008) and elsewhere, the naive approach of constraining every non-terminal to a syntactic constituent severely limits the coverage of the resulting grammar. Therefore, several approaches have been used to improve coverage when using syntactic information. Zollmann and Venugopal 2006 allow rules to be extracted where non-terminals do not exactly span a target constituent. The non-terminals are then labeled with complex labels which amalgamates multiple labels in the span. This increases coverage at the expense of increasing data sparsity as the non-terminal symbol set increases dramatically.

Syntax-Based and Hierarchical Statistical MT

There are a great number of ways in which these two basic methods can be combined together and they will be reviewed below. Basically, what syntactic models do is explicitly to take into account the syntax of the sentences being translated. One simple approach is to limit the phrases learned by a standard PBSMT translation model to only those contiguous sequences of words that additionally correspond to constituents in a syntactic parse tree. However, a total reliance on such syntax-based phrases has been shown to be detrimental to translation quality, as the source-side and target-side tree structures heavily constrain the space of phrase segmentation of a parallel sentence. Noting that the number of phrase pairs extracted from a corpus is reduced by around 80% when they are required to correspond to syntactic constituents, Koehn et al. (2003) observed that many non-constituent phrase pairs that would not be included in a syntax-only model are in fact extremely important to system performance. Since then, researchers have explored effective ways for combining phrase pairs derived from syntax-aware methods with those extracted from more traditional PBSMT (see Xiong et al. 2010a). Briefly stated, the goal is to retain the high level of coverage provided by non-syntactic PBSMT phrases while simultaneously incorporating and exploiting specific syntactic knowledge.

At the same time, it is desirable to include as much syntactic information in the system as possible in order to carry out linguistically motivated reordering, for example: an extended and modified version of the approach of Tinsley et al. (2007), i.e. extracting syntax-based phrase pairs from a large parallel parsed corpus, combining them with PBSMT phrases, and performing joint decoding in a syntax-based MT framework without loss of translation quality. This effectively addresses the low coverage of purely syntactic MT without discarding syntactic information.

A lot of work has focused on combining hierarchical and syntax translation, utilizing the high coverage of hierarchical decoding and the insights that syntactic information can bring. This is done with the aim to balance the generality of using undecorated non-terminals with the specificity of labeled non-terminals. In particular, systems can use syntactic labels from a source language parser to label non-terminal in production rules. However, other source span information, such as chunk tags, can also be used, as will be discussed below.

Researchers have experimented with different methods for combining the hierarchical and syntactic approaches. Syntactic translation rules are used concurrently with a hierarchical phrase rules by training them independently and then using them concurrently to decode sentences.

Another possible method is to use one translation model containing both hierarchical and syntactic rules. Moreover, rules can contain both decorated syntactic non-terminals, and undecorated hierarchical-style non-terminals (in addition, the left-hand-side non-terminal may, or may not be decorated). Improvements may come by using simpler tools: for instance linguistic information coming from shallow parsing techniques – like the chunk tagger (Abney 1991) instead of a full-fledged parser-rule extraction to reduce spurious ambiguity.

Zollmann and Venugopal 2006 etc. overcome the restrictiveness of the syntax-only model by starting with a complete set of phrases as produced by traditional PBSMT heuristics, then annotating the target side of each phrasal entry with the label of the constituent node in the target-side parse tree that subsumes the span. They then introduce new constituent labels to handle the cases where the phrasal entries do not exactly correspond to the syntactic constituents. Liu et al. (2006) also add non-syntactic PBSMT phrases into their tree-to-string translation system.

There has been much effort to improve performance for hierarchical phrase-based machine translation by employing linguistic knowledge. For instance M&R 2008 etc., explore "soft syntactic constraints" on hierarchical phrase model; (Stein et al. 2010) focus on syntactic constraints not only via the constituent parse but also via the dependency parse tree of source or target sentence. (Chiang et al. 2009; Chiang 2010) similarly define many syntactic features including both source and target sides but integrate them into the translation model by MIRA algorithm to optimize their weights.

In particular, M&R 2008 extend a hierarchical PBSMT system with a number of features to prefer or disprefer certain types of syntactic phrases in different contexts. Restructuring the parse trees to ease their restrictiveness is another recent approach: in particular, Chao Wang et al. (2007) binarize source-side parse trees in order to provide phrase pair coverage for phrases that are partially syntactic. Tinsley et al. (2007) showed an improvement over a PBSMT baseline on four tasks in bidirectional German–English and Spanish–English translation by incorporating syntactic phrases derived from parallel trees into the PBSMT translation model. They first word align and extract phrases from a parallel corpus using the open-source Moses PBSMT toolkit (Koehn et al. 2007), which provides a baseline SMT system. Then, both sides of the parallel corpus are parsed with independent automatic parsers, subtrees from the resulting parallel treebank are aligned, and an additional set of phrases (with each phrase corresponding to a syntactic constituent in the parse tree) is extracted. The authors report statistically significant improvements in translation quality, as measured by a variety of automatic metrics, when the two types of phrases are combined in the Moses decoder.

ISI's system obtained best performance on *Ch_En* at NIST 2009. However there are also drawbacks in using this approach and they are all related to the difficulty inherent in producing the needed representation which require language-specific resources (parsers, morphological analysers, etc.). Since parsing is by itself also far from reaching 100% accuracy, the performance of the SBMT system is heavily dependent on parsing quality. Also due to the need to encode additional information to the one represented by simple words, the system will need larger search space and will result in overall costlier processing. Other limitations with the syntax based approaches (such as Quirk et al. 2005; Chiang 2005) are, that they do not offer flexibility for adding linguistically motivated features, and that it is not possible to use morphological factors in the syntax based approaches. In general, the translation quality has shown improvements: in particular, these improvements are due to the more accurate phrase boundary detection. So we may safely say that syntactic phrases are a much more precise representation of translational equivalence, and this is the main reason for adopting such an approach.

Introducing Soft Syntactic Features with Discriminative Classifiers

In the last decade, there has been countless research in soft syntactic features, much of which has led to the improved performance for Hiero. However, it seems that all the syntactic constituent features cannot efficiently work together in the Hiero optimized by MERT. So a more general soft syntactic constraint model has been proposed, based on discriminative classifiers for each constituent type and integrate all of them into the translation model with a unified form. The experimental results show that this method significantly improves the performance on the NIST05 Chinese-to-English translation task.

Soft Syntactic Constraint models (SSC) have been proposed at first by M&R 2008 as heuristic models, while SSC models proposed by Liu et al. (2011) are much more general and based on discriminative classifiers. In this latter paper, they further decompose crossing constituents into three types to contain more syntactic information. For example, similarly to Zollmann and Venugopal 2006, the crossing constituent "NP+" is divided into L\NP, NP/R, and L\NP/R, which means a partial syntactic category NP missing some category to the left, the right and the left and right together, respectively. They are called *general constituent labels (GCL)*. Chiang et al. 2008 introduce heuristic models, that are not sensitive to other features such as boundary word information. However, (Xiong et al. 2006), showed in previous work that these features are helpful for the translation model. On the other hand, uniform combination of all the constituent models may cause a model bias, since some constituent types occur more often than others.

Liu et al. (2011) propose a discriminative soft constraint model for each syntactic constituent type. The underlying idea is to improve the model by integrating it with context information. They consider several classifiers with different accuracy to construct soft constraint models, and they aim to study the effect of the accuracy of the classifiers on the translation performance. Then, they investigate an efficient method to combine all the models to give a unified soft constraint model. Instead of uniformly combining all the models, they introduce a prior distribution for them and combine them with the priority.

The authors propose a unified SSC model based on discriminative classifiers for hierarchical phrase-based translation. Experimental results prove the effectiveness of the method on the NIST05 Chinese-to-English translation task. The experiment shows that the discriminative soft syntactic constraint model achieves better result over the heuristic model of M&R 2008; then, it empirically proves that the more accurate classifier can gain better results when building a sub-model for the translation model. Finally we have an efficient method which integrates all models with respect to general constituent labels into hierarchical phrase translation model and improves its performance.

For different syntactic categories (e.g. NP), M&R 2008 defined some kinds of soft-constraint constituency features (e.g. NP=, NP+, NP_, etc.) for Hiero rules. For instance, if a synchronous rule is used in a derivation, and the span of is a cross constituent "NP+" in the source language parse tree, this rule will get an additional value to the model

score for the case of "NP+". In fact, each of these features can also be viewed as a discrete model with value {0, 1}, i.e. for the case of "NP=" if the span of is exactly "NP", the rule gets a score 1 and 0 otherwise. These constituency features don't distinguish the rules with the same span in the source language. For a training instance corresponding to a rule, inspired by previous work (Zollmann and Venugopal 2006; He et al. 2008; Cui et al. 2010), they design the following features to train SSC models;

Syntactic features, which are the general constituent labels defined in section "Specific Issues in Hybrid MT" for the spans of r and the nonterminal symbols in the source side.
Parts-of-speech (POS) features, which are the POS of the words immediately to the left and right of and those of the boundary words covered by the nonterminal symbols in the source side.
Length features, which are the length of sub-phrases covered by the nonterminal symbols in the source side.

In fact, the models can be extended to include other features, especially those in the target side. In order to compare these models with the work of M&R 2008, they merely introduce several features. They implement a hierarchical phrase-based system as the baseline, similar to Hiero (Chiang 2005), and use XP (M&R 2008) as the comparison system. They use the default setting as Hiero. Word alignment for each sentence pair is obtained as usual. Then, Stanford parser (Klein and Manning 2003) is employed to generate the parse tree for the source side of the data. They acquire about 15.85M training examples among which are 6.81M positive and 9.04M negative examples respectively. There are 88 general constituent labels in all. They employ the open toolkits of MaxEnt and LogReg to train SSC models for each GCL, and construct a linear combination model with them, where the interpolation weight is set to 0.86. They train a 4-gram language model on the Xinhua portion of the English Gigaword corpus using the SRILM Toolkits (Stolcke 2002) with modified Kneser-Ney smoothing (Chen and Goodman 1998). In the experiments, case-sensitive BLEU4 metric (Papineni et al. 2002) measures the translation performances and the statistical significance in BLEU score differences is tested by paired bootstrap re-sampling (Koehn 2004).

Translation Consistency Enforced by Graph-Based Learning

Alexandrescu and Kirchhoff (2009) propose a new graph-based learning algorithm is proposed with structured inputs and outputs to improve consistency in phrase-based statistical machine translation. They define a joint similarity graph over training and test data and use an iterative label propagation procedure to regress a scoring function over the graph. For the purpose of reranking, the resulting scores for unlabeled samples (translation hypotheses) are then combined with standard model scores in a log-linear translation model. From a machine learning perspective, graph-based learning (GBL) is applied to a task with structured inputs and outputs. This is a novel contribution

in itself since previous applications of GBL have focused on predicting categorical labels. The evaluation demonstrates significant improvements over the baseline.

As discussed above, current phrase-based statistical machine translation (SMT) systems commonly operate at the sentence level; each sentence is translated in isolation, even when the test data consists of internally coherent paragraphs or stories, such as news articles. For each sentence, SMT systems choose the translation hypothesis that maximizes a combined log-linear model score, which is computed independently of all other sentences, using globally optimized combination weights. Thus, similar input strings may be translated in very different ways, depending on which component model happens to dominate the combined score for that sentence. A phrase can be translated differently – and wrongly – due to different segmentations and phrase translations chosen by the decoder. Though different choices may be sometimes appropriate, the lack of constraints enforcing translation consistency often leads to suboptimal translation performance. It would be desirable to counter this effect by encouraging similar outputs for similar inputs (under a suitably defined notion of similarity, which may include, for example, a context specification for the phrase/sentence). In machine learning, the idea of forcing the outputs of a statistical learner to vary smoothly with the underlying structure of the inputs has been formalized in the graph-based learning (GBL) framework. In GBL, both labeled (train) and unlabeled (test) data samples are jointly represented as vertices in a graph whose edges encode pairwise similarities between samples. Various learning algorithms can be applied to assign labels to the test samples while ensuring that the classification output varies smoothly along the manifold structure defined by the graph. GBL has been successfully applied to a range of problems in computer vision, computational biology, and natural language processing. However, in most cases, the learning tasks consisted of unstructured classification, where the input was represented by fixed length feature vectors and the output was one of a finite set of discrete labels. In machine translation, by contrast, both inputs and outputs consist of word strings of variable length, and the number of possible outputs is not fixed and practically unlimited.

GBL is an instance of semi-supervised learning, specifically transductive learning. A different form of semi-supervised learning (self-training) has been applied to MT by (Ueffing et al. 2007; Fraser and Marcu 2006). This is the first study to explore a graph-based learning approach. In the machine learning community, work on applying GBL to structured outputs is beginning to emerge. The graph-based learning scheme is used to implement a consistency model for SMT that encourages similar inputs to receive similar outputs. Evaluation on two small-scale translation tasks showed significant improvements of up to 2.6 points in BLEU and 2.8% PER. As the authors report, the approach needs improvements in future work that will include testing different graph construction schemes, in particular better parameter optimization approaches and better string similarity measures; always according to the authors, more gains can be expected when using better domain knowledge in constructing the string kernels. This may include e.g. similarity measures that accommodate POS tags or morphological features, or comparisons of the syntax trees of parsed sentence. The latter could be quite easily incorporated into a string kernel or the related tree kernel similarity measure.

Problems in Combining PBSTM and SBSTM: Rules and Constraints

Galley et al. (2004) create minimal translation rules which can explain a parallel sentence pair but the rules generated are not optimized to produce good translations or coverage in any SMT system. This work was extended and described in (Galley et al. 2006) who create rules composed of smaller, minimal rules, as well as deal with unaligned words. These measures are essential for creating good SMT systems, but again, a parser strictly constrains the rules of syntax.

DeNeefe (2007: 756, 757) proposed the GHKM Galley's – where GHKM is an acronym for the authors names Galley, Hopkins, Knight and Marcu – syntax-based extraction method for learning statistical syntax-based translation rules, presented first in (Galley et al. 2004) and expanded on in (Galley et al. 2006). It is similar to phrase-based extraction in that it extracts rules consistent with given word alignments. A primary difference is the use of syntax trees on the target side, rather than sequences of words. The basic unit of translation is the translation rule, consisting of a sequence of words and variables in the source language, a syntax tree in the target language having words or variables at the leaves, and again a vector of feature values which describe this pair's likelihood. Translation rules can:

- Look like phrase pairs with syntax decoration
- Carry extra contextual constraints
- Be non-constituent phrases
- Contain non-contiguous phrases, effectively "phrases with holes"
- Be purely structural (no words)
- Re-order their children

Decoding with this model produces a tree in the target language, bottom-up, by parsing the foreign string using a CYK parser (Chappelier and Rajman 1998) and a binarized rule set (Zhang et al. 2008). During decoding, features from each translation rule are combined with a language model using a log-linear model to compute the score of the entire translation. The GHKM extractor learns translation rules from an aligned parallel corpus where the target side has been parsed. This corpus is conceptually a list of tuples of 'source sentence, target tree, bi-directional word alignments' which serve as training examples. For each training example, the GHKM extractor computes the set of minimally-sized translation rules that can explain the training example while remaining consistent with the alignments. This is, in effect, a non-overlapping tiling of translation rules over the tree-string pair. If there are no unaligned words in the source sentence, this is a unique set. This set, ordered into a tree of rule applications, is called the derivation tree of the training example. As with ATS (Alignment Template System), translation rules are extracted and counted over the entire training corpus, a count of one for each time they appear in a training example. These counts are used to estimate several features, including maximum likelihood probability features.

To extract all valid tree-to-tree rules, (Liu et al. 2009) extends the famous tree-to-string rule extraction algorithm GHKM (Galley et al. 2004) to their forest-based

tree-to-tree model. However, only with GHKM rules, the rule coverage is very low. As SPMT rules (Marcu et al. 2006) have proven to be a good complement to GHKM (DeNeefe et al. 2007), Zhai et al. also extract full lexicalized SPMT (Marcu et al. 2006: the acronym stands for"Statistical machine translation with syntactified target language phrases") rules to improve the rule coverage.

The tree-to-tree style SPMT algorithm used in their experiments is described as follows:

> … for each phrase pair, traverse the source and target parsing tree bottom up until it finds a node that subsumes the corresponding phrase respectively, then extract a rule whose roots are the nodes just found and the leaf nodes are the phrases.

However, even with GHKM and SPMT rules, the rule coverage is still very low since tree-to-tree models require that both source side and target side of its rule must be a subtree of the parsing tree. With this hard constraint (Liu et al. 2009; Chiang 2010), the model would lose a large amount of bilingual phrases which are very useful to the translation process (DeNeefe et al. 2007). In particular it can be shown that phrase-based models can extract all useful phrase pairs, while string-to-tree and tree-to-string model can only extract part of them because of the one-side subtree constraint. Further, with the rigid ***both-side subtree constraint*****,** the rule space of tree-to-tree model is the narrowest, accounting only for at most 8.45% of all phrase pairs. Hence, learning to enlarge the rule coverage is the challenge for tree-to-tree models.

In the decoding process, the procedure traverses the source parsing tree in a bottom up fashion and tries to translate the subtree rooted at the current node. If the employed rule is full lexicalized, *candidate translations* are generated directly. Otherwise new candidate translations are created by combining target terminals of the rule and candidate translations of the corresponding descendant nodes of the current node. Root node of the parsing tree will be the last visited node and the best translation is chosen as usual from its best candidate translations. Broadly, tree-to-tree based decoding is node-based, i.e., only the source spans governed by tree nodes can be translated as a unit. These spans are called ***translation spans***. During decoding, translation spans are used for translation, while other spans are ignored completely even if they include better translations. Thus this rigid constraint (they call it ***node constraint)*** will exclude many good translations. Zhai et al. (2011) use the Chinese part of the FBIS corpus as a test set: in their statistics, there are in total of 14.68M effective translation spans in the corpus. However, only 44.6% (6.54M spans) of them are governed by tree nodes. This low proportion would definitely lead to an exceptionally narrow search space for tree-to-tree model and a poor translation quality.

In addition, the model is also heavily affected by the ***exact matching constraint*** which means only the rules completely matching part of the source tree structure can be used for decoding. Since parsing errors are very common with automatic parsers, the mismatch is not rare. Moreover, the large and flat structures which have a close relation with reordering are also hard to match exactly. Thus with such constraint, many rules cannot be employed during decoding even if by the model extracts them and the search space is necessarily decreased.

In order to resolve the constraints, two simple but very effective approaches are proposed: (1) integrating bilingual phrases to improve the rule coverage problem; (2) binarizing the bilingual parsing trees to relieve the rigid syntactic constraints. Other systems using transducers with MLE probabilities may also benefit from additional reordering models (more on this topic below).

Huang et al. (2010) decorate the syntax structure into the non-terminal in hierarchical rules as a feature vector. During decoding time, they calculate the similarity between the syntax of the source side and the rules used to derive translations, and then they add the similarity measure to translation model as an additional feature. They don't directly use the syntax knowledge to calculate the additional feature score, but use it to derive a latent syntactic distribution. He et al. (2008) and Cui et al. (2010) employ the syntax knowledge as some of the features to construct rule selection models. When training discriminative models training examples are derived from the rule extraction or from the formal bilingual parsing derivation forest of the training data. Their strong results reinforce the claim that discriminative models are useful in building the sub-model in translation.

Huang and Chiang (2007) use parse information of the source language, and their production rules consist of source tree fragments and target languages strings. During decoding, a packed forest of the source sentence is used as input, and the production rule tree fragments are applied to the packed forest. Liu et al. (2009) use joint decoding with a hierarchical and tree-to-string model and find that translation performance increases for a Chinese-English task.

Others have sought to add soft linguistic constraints to hierarchical models using addition feature functions, such as M&R 2008 who add feature functions to penalize or reward non-terminals which cross constituent boundaries of the source sentence. Shen et al. (2009) discuss soft syntax constraints and context features in a dependency tree translation model. The POS tag of the target head word is used as a soft constraint when applying rules. Also, a source context language model and a dependency language model are used as features. Most SMT systems use the Viterbi approximation whereby the derivations in the log-linear model are not marginalized, but the maximum derivation is returned. String-to-tree models build on this so that the most probable derivation, including syntactic labels, is assumed to be the most probable translation. This fragments the derivation probability and further partitions the search space, leading to pruning errors. Venugopal et al. (2009) attempts to address this by efficiently estimating the score over an equivalent unlabeled derivation from a target syntax model. Ambati and Lavie (2008) and Ambati et al. (2009) note that tree-to-tree often underperforms models with parse tree only on one side due to the non-isomorphic structure of languages. This motivates the creation of an isomorphic backbone into the target parse tree, while leaving the source parse unchanged.

Hoang and Koehn (2010), present a new translation model that includes undecorated hierarchical-style phrase rules, decorated source-syntax rules, and partially decorated rules. Results show an increase in translation performance of up to 0.8% BLEU for German–English translation when trained on the news-commentary corpus, using syntactic annotation from a source language parser. Also experimenting with annotation from shallow taggers may increase BLEU scores.

This continues earlier work in (Chiang 2005) but they see gains when finer grain feature functions are used. The weights for feature function is tuned in batches due to the deficiency of MERT when presented with many features. Chiang et al. (2008) rectified this deficiency by using the MIRA to tune all feature function weights in combination. However, the translation model continues to be hierarchical. Chiang et al. (2009) added thousands of linguistically-motivated features to hierarchical and syntax systems, However, the source syntax features are derived from the research above. The translation model remains constant but the parameterization changes.

Syntax Based SMT and Fuzzy Methods

Tree-to-tree translation models suffer from unsatisfactory performance due to the limitations both in rule extraction and decoding procedure, and in several rigid syntactic constraints that severely hamper these models. These constraints include: the both-side subtree constraint in rule extraction, the node constraint and the exact matching constraint in decoding. Zhai et al. (2011) propose two simple but effective approaches to overcome the constraints: utilizing fuzzy matching and category translating to integrate bilingual phrases and using head-out binarization to binarize the bilingual parsing trees. Their experiments show that the proposed approaches can significantly improve the performance of tree-to-tree system and outperform the state-of-the-art phrase-based system Moses.

Two main directions have emerged to overcome the limitations discussed above. One is to loose the syntactic constraints. (Zhang et al. 2008) proposes a *tree-sequence based tree-to-tree model* that represents rules with tree sequences and takes all spans as translation spans. This method resolves the both-side subtree constraint and the node constraint thoroughly, but it neglects the bad influence of the exact matching constraint. Furthermore, it is obvious that each bilingual phrase would multiply into many tree sequence rules with different structures, which definitely leads to serious rule expansion to increase the decoding burden. In the other direction, more information is introduced into the model. (Liu et al. 2009) substitutes one-best tree with packed forest for tree-to-tree model which can compactly encode many parses and successfully relieve the constraints. But even with packed forest, the rule coverage is still very low. The two directions have proven to outperform their conventional counterparts significantly. However, whether tree sequence or packed forest, they are all complicated to deal with in decoding stage, and furthermore, they both need to modify the conventional tree-to-tree model. Thus they must heavily adjust the original decoding algorithm to cater for the corresponding changes.

To improve the conventional tree-to-tree model the authors propose integrating bilingual phrases and binarizing the bilingual parsing trees. (Liu et al. 2006) and (Mi et al. 2008) utilize bilingual phrases to improve tree-to-string and forest-to-string model. Other authors integrate bilingual phrases into tree-to-tree model to resolve the problem of poor coverage of rules. Of the two, this model is the more

difficult since it must provide syntactic structures for both the source and target phrases to serve the decoding process of the model.

In traditional tree-to-tree based decoding, source side of the rule is employed to match the source parsing tree exactly. Thus if we want to use a source phrase, theoretically we must decorate it with the corresponding syntax structure like the tree-sequence based model. However, it has been shown that exact match would do harm to the translation quality. Thus instead of syntax structures, source phrases are decorated with syntactic categories which are necessary and effective for translation (Zhang et al. 2011). When decoding with these source phrases, the system ignores the internal structure of the subtree for translation and only matches the rule's category with root node of the subtree along with the matching between leaf nodes. Normally, if the system tries an exact match, a given rule may not be employed in case of mismatch between categories of rule and tree structure. Hence, to maximize the capacities of the source phrases, the fuzzy matching method can be employed which has been successfully employed in hierarchical phrase-based model (Huang et al. 2010) and string-to-tree model (Zhang et al. 2011) to match categories. With fuzzy matching method, Zhai et al. (2011) represent each SAMT-style syntactic category with a real-valued vector $F_{(c)}$ using latent syntactic distribution. That is to say, they transform an original source phrase by decorating it with a SAMT-style syntactic category and a corresponding real-valued vector. During decoding, they consider all possible source phrases and compute the similarity scores between categories of phrases and head nodes of the current translated structure. Then the similarity score will serve as a good feature *(similarity score feature)* incorporated into the model and will let it learn how to respect the source phrases.

Combining PBSTM and SBSTM but Then Syntax-Prioritizing

A key concern in building syntax-based machine translation systems is how to improve coverage by incorporating more traditional phrase-based SMT phrase pairs that do not correspond to syntactic constituents. Improved precision due to the inclusion of syntactic phrases can be seen by examining a translation example and the phrasal chunks chosen which exist in the baseline PBSMT phrase table, but do not make it into the top-best translation in the PBSMT-only scenario because of its high ambiguity factor. Hanneman and Lavie (2009) propose an approach which is structurally similar to that of Tinsley et al. (2007), extended or modified in a number of key ways. At first, they extract both non-syntactic PBSMT and syntax-driven phrases from a parallel corpus that is two orders of magnitude larger. Then, they apply a different algorithm for subtree alignment, proposed by Lavie et al. (2008), which proceeds bottom-up from existing statistical word alignments, rather than inducing them top-down from lexical alignment probabilities. In addition to combining straightforwardly syntax-derived phrases with traditional PBSMT phrases, they propose a new combination technique that removes PBSMT phrases whose source-language strings are already covered by a syntax-derived phrase. This new

syntax-prioritized technique results in a 61% reduction in the size of the combined phrase table with only a minimal decrease in automatic translation metric scores. Finally, and crucially, they carry out the joint decoding over both syntactic and non-syntactic phrase pairs in a syntax-aware MT system, which allows a syntactic grammar to be put in place on top of the phrase pairs to carry out linguistically motivated reordering, hierarchical decoding, and other operations.

A small number of grammar rules are then used to correct the structure of constituents which require some reordering in the sentence. After inspecting the output of the test set they find that the grammar is 97% accurate in its applications, making helpful reordering changes 88% of the time.

The statistical transfer ("Stat-XFER") framework (Lavie 2008; and recent extension by Ambati and Lavie 2008) is the base MT system used for an experiment that we report here. It is similar to what we already discussed under section "Combining Syntax, EBMT and a Transfer Approach to MT" making exception for the stochastic Example-Based approach. The core of the framework is a transfer engine using two language-pair-dependent resources: a grammar of weighted synchronous context-free rules, and a probabilistic bilingual lexicon. Once the resources have been provided, the Stat-XFER framework carries out translation in a two-stage process, first applying the lexicon and grammar to parse synchronously an input sentence, then running a monotonic decoder over the resulting lattice of scored translation pieces assembled during parsing to produce a final string output (see Dyer et al. 2008). Reordering is applied only in the first stage, driven by the syntactic grammar; the second-stage monotonic decoder only assembles translation fragments into complete hypotheses. Each Stat-XFER bilingual lexicon entry has a synchronous context-free grammar (SCFG) expression of the source- and target-language production rules. The SCFG backbone may include lexicalized items, as well as non-terminals and pre-terminals from the grammar. Constituent alignment information specifies one-to-one correspondences between source-language and target-language constituents on the right-hand side of the SCFG rule. Rule scores for grammar rules, if they are learned from data, are calculated in the same way as the scores for lexical entries. The grammar and lexicon are extracted from a large parallel corpus that has been statistically word-aligned and independently parsed on both sides with automatic parsers. Word-level entries for the bilingual lexicon are directly taken from word alignments; corresponding syntactic categories for the left-hand side of the SCFG rules are obtained from the preterminal nodes of the parse trees. Phrase-level entries for the lexicon are based on node-to-node alignments in the parallel parse trees. In the straightforward "tree-to-tree" scenario, a given node ns in one parse tree S will be aligned to a node nt in the other parse tree T if the words in the yield of ns are all either aligned to words within the yield of nt or have no alignment at all. If there are multiple nodes nt satisfying this constraint, the node in the tree closest to the leaves is selected. Each aligned node pair (ns,nt) produces a phrase-level entry in the lexicon, where the left-hand sides of the SCFG rule are the labels of ns and nt, and the right-hand sides are the yields of those two nodes in their respective trees. In the expanded "tree-to-tree-string" configuration, if no suitable node nt exists, a new node n′s is introduced into T as a projection of ns, spanning the yield of the words in T aligned to the yield of ns.

Conceptually, they take the opposite approach to that of Tinsley et al. (2007) by adding traditional PBSMT phrases into a syntax-based MT system rather than the other way around. They begin by running steps 3 through 5 of the Moses training script (Koehn et al. 2007), which results in a list of phrase pair instances for the same word-aligned corpus to which they applied the syntax-based extraction methods. Given the two sets of phrases, they explore two methods of combining them: direct combination and syntax-prioritized combination.

- **Direct Combination.** Following the method of Tinsley et al. (2007), they directly combine the counts of observed syntax-based phrase pairs with the counts of observed PBSMT phrase pairs. This results in a modified probability model in which a higher likelihood is moved onto syntactic phrase pairs that were also extractable using traditional PBSMT heuristics. It also allows either extraction mechanism to introduce new entries into the combined phrase table that were not extracted by the other, thus permitting the system to take full advantage of complementary information provided by PBSMT phrases that do not correspond to syntactic constituents.
- **Syntax-Prioritized Combination.** Under this method, they take advantage of the fact that syntax-based phrase pairs are likely to be more precise translational equivalences than traditional PBSMT phrase pairs, since constituent boundaries are taken into account during phrase extraction. PBSMT phrases whose source-side strings are already covered by an entry from the syntactic phrase table are removed; the remaining PBSMT phrases are combined as in the direct combination method above. The effect on the overall system is to trust the syntactic phrase pairs in the cases where they exist, supplementing with PBSMT phrase pairs for non-constituents.

Syntax MT and Dependency Structures

Hoang and Koehn (2010) present an experiment which shows how both hierarchical and syntax-based SMT can be used fruitfully to improve the performance of a system. Japanese and Chinese are the two languages mostly involved in experimenting with syntax-based MT, in particular due to structural differences between the two languages and English. Nakazawa and Kurohashi (2011) introduce a tree-based reordering model which models word or phrase dependency relations in dependency tree structures of source and target languages. They propose a phrase alignment method which models word or phrase dependency relations in dependency tree structures of source and target languages. For a pair of correspondences which has a parent–child relation on one side, the dependency relation on the other side is defined as the relation between the two correspondences. It is a kind of tree-based reordering model, and can capture non-local reorderings which sequential word-based models often cannot handle properly. The model is also capable of estimating phrase correspondences automatically without heuristic rules. The model is trained in two steps: Step 1 estimates word translation probabilities, and Step 2 estimates

phrase translation probabilities and dependency relation probabilities. Both Step 1 and Step 2 are performed iteratively by the EM algorithm. During the Step 2 iterations, word correspondences are grown into phrase correspondences.

Experimental results of alignment show that the model could achieve F-measure 1.7 points higher than the conventional word alignment model with symmetrization algorithms.

The authors consider that there are two important needs in aligning parallel sentences written in very different languages such as Japanese and English. One is to adopt structural or dependency analysis into the alignment process to overcome the difference in word order. The other is that the method needs to have the capability of generating phrase correspondences, that is, one-to-many or many-to-many word correspondences.

Nakazawa and Kurohashi (2008) also proposed a model focusing on the dependency relations. Their model has the constraint that content words can only correspond to content words on the other side, and the same applies for function words. This sometimes leads to an incorrect alignment. Thus they have removed this constraint to make more flexible alignments possible. Moreover, in their model, some function words are brought together, and thus they cannot handle the situation where each function word corresponds to a different part. The smallest unit of our model is a single word, which should solve this problem.

Chang et al. (2011) note that structural differences between Chinese and English are a major factor in the difficulty of machine translation from Chinese to English. The wide variety of such Chinese-English differences include the ordering of head nouns and relative clauses, and the ordering of prepositional phrases and the heads they modify. Previous studies have shown that using syntactic structures from the source side can help MT performance on these constructions. Most of the previous syntactic MT work has used phrase structure parses in various ways, either by doing syntax-directed translation to translate directly parse trees into strings in the target language (Huang et al. 2006), or by using source-side parses to preprocess the source sentences (Wang et al. 2007). One intuitive solution for using syntax is to capture different Chinese structures that might have the same meaning and hence the same translation in English. But it turns out that phrase structure (and linear order) are not sufficient to capture this meaning relation. Two sentences with the same meaning can have different phrase structures and linear orders. They propose to use *typed dependency* parses instead of phrase structure parses. Typed dependency parses give information about grammatical relations between words, instead of constituency information. They capture syntactic relations, such as *nsubj* (nominal subject) and *dobj* (direct object),, but also encode semantic information such as in the *loc* (localizer) relation. This suggests that this kind of semantic and syntactic representation could have more benefit than phrase structure parses. Chinese typed dependencies are automatically extracted from phrase structure parses. In English, this kind of typed dependencies has been introduced by de Marneffe and Manning (2008) and de Marneffe et al. (2006). Using typed dependencies, it is easier to read out relations between words, and thus the typed dependencies have been used in meaning extraction tasks. Features over the Chinese typed dependencies are used in a phrase-based MT system when deciding

whether one chunk of Chinese words (MT system statistical phrase) should appear before or after another. To achieve this, a discriminative phrase orientation classifier is trained following the work by Zens and Ney (2006), and the system uses grammatical relations between words as extra features to build the classifier. Then the phrase orientation classifier is applied as a feature in a phrase-based MT system to help reordering. Basic reordering models in phrase-based systems use linear distance as the cost for phrase movements (Koehn et al. 2003). The disadvantage of these models is their insensitivity to the content of the words or phrases. More recent work (Tillman 2004; Och & Ney 2004; Koehn et al. 2007) has introduced lexicalized reordering models which estimate reordering probabilities conditioned on the actual phrases. Lexicalized reordering models have brought significant gains over the baseline reordering models, but one concern is that data sparseness can make estimation less reliable. Zens and Ney (2006) proposed a discriminatively trained phrase orientation model and evaluated its performance as a classifier and when plugged into a phrase-based MT system. Their framework allows us easily to add in extra features. Therefore it is used as a testbed to see if features from Chinese typed dependency structures can effectively be used to help reordering in MT.

The target language (English) translation is built from left to right. The phrase orientation classifier predicts the start position of the next phrase in the source sentence. They use the simplest class definition and group the start positions into two classes: one class for a position to the left of the previous phrase (reversed) and one for a position to the right (ordered). The basic feature functions are similar to what Zens and Ney (2006) used in their MT experiments. The basic binary features are source words within a window of size 3 around the current source position *j,* and target words within a window of size 3 around the current target position *i.* Yhe classifier experiments in Zens and Ney (2006) also uses word classes to introduce generalization capabilities. In the MT setting it's harder to incorporate the part-of-speech information on the target language. Zens and Ney (2006) also exclude word class information in the MT experiments. In the work they also use word features as basic features for the classification experiments. Assuming the Chinese sentence to translate has been parsed and grammatical relations in the sentence have been extracted, the path between the two words annotated by the grammatical relations is used. This feature helps the model learn the relation between the two chunks of Chinese words. The feature is defined as follows: for two words at positions p and q in the Chinese sentence ($p<q$), find the shortest path in the typed dependency parse from p to q, concatenate all the relations on the path and use that as a feature.

Cherry and Lin (2003) proposed a model which uses a source side dependency tree structure and constructs a discriminative model. However, there is the defect that its alignment unit is a word, so it can only find one-to-one alignments. On the contrary, when aligning very different language pairs, the most important need is the capability of generating both one-to-many and many-to-many correspondences.

Venkatapathy et al. (2010) propose an English-Hindi dependency-based statistical system that uses discriminative techniques to train its parameters. The use of syntax (dependency tree) allowed them to address the large word-reorderings between English and Hindi. And, discriminative training allows us to use rich feature sets, including linguistic features that are useful in the machine translation task.

Morphological decomposition is useful where there is very limited parallel corpora available, and breaking words into smaller units helps in reducing sparsity. In order to handle phenomena, such as long-distance word agreement to achieve accurate generation of target language words, the inter-dependence between the factors of syntactically related words needs to be modeled effectively.

Knowledge-Based MT Systems

Our focus in this section will be knowledge-based systems, i.e. systems which are a combination of both syntax and semantic knowledge to inform statistical models and learning. In particular, I assume that both syntax and semantics should also inform automatic evaluation in order to improve precision. Semantics in this case refers to ontologies like SUMO or WORDNET, but then, in order to be effective, should also include some Word-Sense Disambiguation or at least semantic similarity processing step. Other recent procedures for assessing – and evaluating – semantic similarity are based on Text Entailment techniques, but are less frequently used. Taxonomies and ontologies are data structures that organise conceptual information by establishing relations among concepts, hierarchical and partitive relations being the most important ones. One of the first idea was that of using a multilingual ontology as an interlingua (Hovy and Nirenburg 1992; Hovy 1998; Hovy et al. 2006; Philpot et al. 2010). Nowadays, ontologies have a wide range of uses in many domains, for example, finance (International Accounting Standards Board 2007), bio-medicine (Collier et al. 2008; Ashburner et al. 2000) and libraries (Mischo 1982). These resources normally attach labels in natural language to the concepts and relations that define their structure, and these labels can be used for a number of purposes, such as providing user interface localization (McCrae et al. 2011), multilingual data access (Declerck et al. 2010), information extraction (Müller et al. 2004) and natural language generation (Bontcheva 2005). Applications that use such ontologies and taxonomies will require translation of the natural language descriptions associated with them in order to adapt these methods to new languages. Currently, there has been some work on the idea of multilinguality in ontologies such as EuroWordNet (Vossen 1998), bilingual WordNet, or BOW (Huang et al. 2010), and in the context of ontology localisation, such as Espinoza et al. (2008) and (2009), Cimiano et al. (2010), Fu et al. (2010) and Navigli and Penzetto (2010). Current work in machine translation has shown that word sense disambiguation can play an important role by using the surrounding words as context to disambiguate terms (Carpuat and Wu 2007; Apidianaki 2009).

One of the most interesting hypothesis is the one underlying interlingua RBMT systems. It uses an abstract intermediate semantic/logical representation to be used for translating into any target language. This hypothesis is converted into a SMT-viable alternative in which predicate-argument structures of both source and target language bitexts are used to bootstrap the SMT alignment module. This is what Wu and Palmer (2011) propose with the aim to abstract away from language specific syntactic variation and provide a more robust, semantically coherent alignment across sentences. As

the authors comment, a number of previous attempts had been made to either align deep syntactic/semantic lemmatized representations (as Marecek 2009a, b) did for English/Czech parallel corpus alignment); or to introduce semantic roles and syntax based argument similarity to project English Framenet to German, where however only the source was annotated. Choi et al. (2009) and Wu et al. (2010) enhanced Chinese-English verb alignments using parallel PropBanks. However there was no explicit argument mapping between the aligned predicate-argument structures.

HPBMT with Semantic Role Labeling

Recently there has been increased attention on using semantic information in machine translation. Pighin and Màrquez (2011) present a model for the inclusion of semantic role annotations in the framework of confidence estimation for machine translation. The model has several interesting properties, most notably: (1) it only requires a linguistic processor on the (generally well-formed) source side of the translation; (2) it does not directly rely on properties of the translation model (hence, it can be applied beyond phrase-based systems). These features make it potentially appealing for system ranking, translation re-ranking and user feedback evaluation. Preliminary experiments in pairwise hypothesis ranking on five confidence estimation benchmarks show that the model has the potential to capture salient aspects of translation quality.

Liu and Gildea (2008, 2010) proposed using Semantic Role Labels (SRL) in their tree-to-string machine translation system and demonstrated improvement over conventional tree-to-string methods. Wu and Fung (2009) developed a framework to reorder the output using information from both the source and the target SRL labels, and their approach uses the target side SRL information in addition to a Hierarchical Phrase-based Machine Translation framework. The proposed method extracts initial phrases with two different heuristics. The first heuristic is used to extract rules that have a general left-hand-side (LHS) non-terminal tag X, i.e., Hiero rules. The second will extract phrases that contain information of SRL structures. The predicate and arguments that the phrase covers will be represented in the LHS non-terminal tags. After that, they obtain rules from the initial phrases in the same way as the Hiero extraction algorithm, which replaces nesting phrases with their corresponding non-terminals. By applying this scheme, rules will contain SRL information, without sacrificing the coverage of rules. Such rules are called SRL-aware SCFG rules. During decoding, both the conventional Hiero-style SCFG rules with general tag X and SRL-aware SCFG rules are used in a synchronous Chart Parsing algorithm. Special conversion rules are introduced to ensure that whenever SRL-aware SCFG rules are used in the derivation, a complete predicate-argument structure is built. Gao and Vogel (2011) propose of using Semantic Role Labels to assist hierarchical phrase-based MT. They present a novel approach of utilizing Semantic Role Labeling (SRL) information to improve Hierarchical Phrase-based Machine Translation, by proposing an algorithm to extract SRL-aware Synchronous Context-Free Grammar (SCFG) rules. Conventional Hiero-style SCFG rules are extracted in the same

framework. Special conversion rules are applied to ensure that when SRL-aware SCFG rules are used in derivation, the decoder only generates hypotheses with complete semantic structures. They then perform machine translation experiments using nine different Chinese-English test-sets. The approach achieved an average BLEU score improvement of 0.49 as well as 1.21 point reduction in TER.

When dealing with formalisms such as semantic role labeling, the coverage problem is also critical., so it is important to follow Chiang's (2007) observation to use SRL labels to augment the extraction of SCFG rules. The formalism provides additional information and more rules instead of restrictions that remove existing rules. This preserves the coverage of rules.

Multiword Units in Dependency Structure

In Hwidong Na and Jong-Hyeok Lee (2011), another important contribution comes from the use of multiword units which had already been proposed in the computer assisted MT scenario by Wehrli et al. (2009). Here on the contrary, the translation requires non-isomorphic transformation from the source to the target. However, learning multi-word units (MWUs) can reduce non-isomorphism. They present a novel way of representing sentence structure based on MWUs, which are not necessarily continuous word sequences. The proposed method builds a simpler structure of MWUs than words using words as vertices of a dependency structure. Unlike previous studies, they collect many alternative structures in a packed forest. As an application of the proposed method, they extract translation rules in the form of a source MWU-forest to the target string, and verify the rule coverage empirically. On the same subject see also Carpuat and Diab 2010; Lambert and Banchs 2005; Ren et al. 2009.

Ontologies and Taxonomies

McCrae et al. (2011) widely use ontologies and taxonomies to organize concepts providing the basis for activities such as indexing and as background knowledge for NLP tasks. As such, translation of these resources would prove useful to adapt these systems to new languages. However, they show that the nature of these resources is significantly different from the "free-text" paradigm used to train most statistical machine translation systems. In particular, significant differences in the linguistic nature of these resources can be seen and such resources have rich additional semantics. As a result of these linguistic differences, standard SMT methods, in particular evaluation metrics, can produce poor performance. Leveraging these semantics for translation can be approach in three ways: by adapting the translation system to the domain of the resource; by examining if semantics can help to predict the syntactic structure used in translation; and by evaluating if existing translated taxonomies can be used to disambiguate translations. Results from these experiments shed light on

the degree of success that may be achieved with each approach. Rather than looking for exact or partial translations in other similar resources such as bilingual lexica, in the paper an adequate translation is presented using statistical machine translation approaches that also utilise the semantic information beyond the label or term describing the concept, that is relations among the concepts in the ontology, as well as the attributes or properties that describe concepts.

Latent Semantic Indexing

It is evident that the main-stream statistical machine translation is unable to tackle source-context information in a reliable way has been already recognized as a major drawback of the statistical approach, whereas (Carl and Way 2003) have proven the use of source-context information has been proven to be effective in the case of example-based machine translation. In this regard, (Carpuat and Wu 2007, 2008; Haque et al. 2009; España-Bonet et al. 2009; Banchs and Costa-jussà 2010) have already reported attempts to incorporate source-context information into the phrase-based machine translation framework. However, no transcendental improvements in performance have been achieved or, at least, reported yet.

Rafael E. Banchs & M.R. Costa-jussà (2011) proposed and evaluated an approach that uses a semantic feature for statistical machine translation, based on Latent Semantic Indexing. The objective of the proposed feature is to account for the degree of similarity between a given input sentence and each individual sentence in the training dataset. This similarity is computed in a reduced vector-space constructed by means of the Latent Semantic Indexing decomposition. The computed similarity values are used as an additional feature in the log-linear model combination approach to statistical machine translation. In the implementation, the proposed feature is dynamically adjusted for each translation unit in the translation table according to the current input sentence to be translated. This model aims to favor those translation units that were extracted from training sentences that are semantically related to the current input sentence being translated. Experimental results on a Spanish-to-English translation task on the Bible corpus demonstrate a significant improvement on translation quality with respect to a baseline system.

Crucial semantic problems are dealt with in a recent paper by Baker et al. 2012, on semantic issues like modality and negation which are relevant for SMT or what they call Semantically Informed Syntactic MT.

Evaluation Methods and Tools

To comment on this topic I will refer to a paper by Forcada et al. (2011b) – but see also Daelemans and Hoste 2009 – who extensively presents and experiments with evaluation metrics. One of the most widely used automatic MT evaluation metrics

is BLEU; then we have the NIST evaluation metric, the GTM metric based on precision and recall. Some of the metrics presented are language specific, and they are: METEOR, METEOR-NEXT, TER-plus and DCU-LFG. These metrics need specific resources and tools which are at present only available for English – see Kirchhoff et al. 2007 for semi-automatic evaluation.

BLEU (Papineni et al. 2002) – the most widely used automatic MT evaluation metrics – is a string-based metric which has come to represent something of a de facto standard in the last few years. This is not surprising given that today most MT research and development efforts are concentrated on statistical approaches; BLEU's critics argue that it tends to favour statistical systems over rule-based ones (Callison-Burch et al. 2006). Using BLEU is fast and intuitive, but while this metric has been shown to produce good correlations with human judgment at the document level (Papineni et al. 2002), especially when a large number of reference translations are available, correlation at sentence level is generally low. BLEU measures n-gram precision and the score is between 0 and 1. N-grams considered are any linear sequence of words up to length 4 (BLEU4), to be found in actual output and in reference translation. There is a brevity penalty in that single word match is just not counted if it never appears alone, like for instance the word "the"; and the same portion of text can't be used. BLEU is not sensitive to global syntactic structure; it doesn't care if the wrong translation is a function word rather than a content words or a proper name (input source sentences are all lower-cased and upper-case words are no longer visible). Human translation scored by BLEU typically falls around 60% – rather than 100% due to translator variations – and the best Chinese or Arabic translations into English may reach the same value (as reported in Ravi and Knight 2010).

The NIST evaluation metric (Doddington 2002) is also string-based, and gives more weight in the evaluation to less frequent n-grams. While this metric has a strong bias in favour of statistical systems, it provides better adequacy correlation than BLEU (Callison-Burch et al. 2006).

The GTM metric (Turian et al. 2003) is based on standard measures adopted in other NLP applications (precision, recall and F-measure), which makes its use rather straightforward for NLP practitioners. It focuses on unigrams and rewards sequences of correct unigrams, applying moderate penalties for incorrect word order.

METEOR (Banerjee and Lavie 2005; Lavie and Agarwal 2007) uses stemming and synonymy relations to provide a more fine-grained evaluation at the lexical level, which reduces its bias towards statistical systems. One drawback of this metric is that it is language-dependent since it requires a stemmer and WordNet.3. It can currently be applied in full only to English, and partly to French, Spanish and Czech, due to the limited availability of synonymy and paraphrase modules. METEOR-NEXT (Denkowski and Lavie 2010) is an updated version of the same metric.

The TER metric (Snover et al. 2006) adopts a different approach, in that it computes the number of substitutions, insertions, deletions and shifts that are required to modify the output translation so that it completely matches the reference translation(s). Its results are affected less by the number of reference translations than is the case for BLEU. Also, the rationale behind this evaluation metric is quite simple to understand for people who are not MT experts, as it provides an estima-

tion of the amount of post-editing effort needed by an end-user. Another metric based on error rates which preceded TER is WER (Nießen et al. 2000). WER and its extension mWER (Nießen et al. 2000) have been omitted from the experiments reported here as they seem to have been superseded by more recent metrics.

TER-plus (Snover et al. 2009) is an extension of TER using phrasal substitutions relying on automatically generated paraphrases, stemming, synonyms and relaxed shifting constraints. This metric is language-dependent and requires WordNet. It has been shown to have the highest average rank in terms of Pearson and Spearman correlation (Przybocki et al. 2008).

The DCU-LFG metric (Owczarzak et al. 2007) exploits LFG dependencies and has only a moderate bias towards statistical systems. It requires a dependency parser. Xiong et al. 2010b use linguistic feature to detect errors.

It should be noted that among the above measures, METEOR, METEOR-NEXT, TER-plus and DCU-LFG can only be used for English as a target language at the present time, given the language-specific resources that they require.

A Translation Example: Comparisons and Comments

Just to show how syntax, morphology and semantics may play an important role, we will use an example (Wilks/Zampolli 1994:592) from one of the many papers that Yorick Wilks has published on the subject (see Wilks 2009). The example is interesting in that it introduces the need to take care of agreement in discontinuous constituents. We will use common online translation systems (a RBMT Systrans and a SMT Google), will compare translations into three common European languages, and will comment on the errors produced:

(1) The soldiers fired at the women and I saw several fall
Google: (Ita) I soldati hanno sparato alle donne e ho visto cadere molti
(Fre) Les soldats ont tiré sur les femmes et j'ai vu la chute de plusieurs
(Germ) Die Soldaten schossen auf die Frauen, und ich sah mehrere Sturz
---> I soldati hanno sparato contro le donne, e ho visto cadere molti
Systran: (Ita) I soldati fatti fuoco contro le donne e me hanno veduto parecchio caduta
(Fre) Les soldats mis le feu aux femmes et à moi ont vu plusieurs chute
(Germ) Die Soldaten, die an den Frauen und an mir gefeuert wurden, sahen einiges Fall

On the whole, Google produces acceptable translations even though agreement is wrong both in Italian and French. In addition, German translation introduces a noun instead of the infinitival – Sturz translates the base verb form "fall" treating it as a noun. The treatment of the complement of "fired" (at the women) are all fine in the three languages, which we certainly regard as an achievement possible thanks to statistics: both the preposition and case are fine. If we look at Systran's translation on the contrary, we see an attempt to control gender in Italian: "caduta" is a feminine singular, but "parecchio" is masculine singular. So it would seem that there is no provision for the treatment of Number (both should have been plural). Another mistake is the presence of "me" in front of "hanno veduto", which is not only wrong –

"vedere"/see is not like "piacere", a psych verb that turns the deep experiencer subject into a dative. In fact "me" is accusative and as such it cannot be used in the subject position of any verb unless it is followed by another clitic that is in the accusative form as in "me lo"/to me it. The mistake is clearly due to the wrong phrase produced by joining "at the women and I" as if they were bound to the same preposition "at".

The auxiliary form is right but the past participle "veduto" is an archaic or stylistically marked version of "visto" that translates the simple past of "see". The question is that the main clause does not have a tensed main verb anymore: "fatti" is an absolute participial clause and translates "fired" as a past participle and not as a simple past. This ambiguity is quite common in English: in fact almost all verbs – with the exception of those irregular forms that are different in the two tenses, like "went/gone" – are ambiguous and require a disambiguation procedure in the parser to tell one tense from the other. The mistake is clearly related to the lack of statistical measures associated to the choice. Also French is semantically wrong: "mis le feu" does not really fit into the translation required here, it translates the other meaning of "fire", BURN. It is difficult to understand the semantic choice here, because you don't currently BURN WOMEN very easily, even though that might have happened in the past. Actually in the Middle-Ages when witches were around, many women were set on fire on a pyre. As to the conjoined sentence, we see again the same mistake of using a dative "à moi" in French and "an mir" in German, rather than simply introducing a nominative pronoun, like "moi" and "ich". Then the quantifier "several" is again translated without agreement in both French and German: however French "plusieurs" captures Number and the German "einiges" is wrong both in agreement and in meaning – it translates "some". In fact, the German translation is primarily wrong because it turns the two conjoined sentences as if they were headless relative clauses – which is possible in English but not in German – parsing the constituents "fired at the women and I" as if they were a well-formed structure. This is apparent from the introduction of two commas, at the beginning and at the end of the conjunct "…, die an den Frauen und an mir gefeuert wurden," and by the use of passive which is clearly nonsensical given the presence of a nominative Agent "die Soldaten".

So eventually, Systran has produced a far worse result than Google, which by making use of its enormous terabyte of parallel texts, has shown the power of SMT. More examples follow below.

MT for the Future

New Statistical Methods and a Comprehensive Translation Model

We assume that the right direction for MT of the future is to incorporate both syntax and semantics in its statistics. We can do this in these two ways:

A. A first way would be the one proposed by the LOGON project (Oepen et al. 2004, 2005; Oepen and Lønning 2006). It increases the role of NLP tools and

leaves mainly to statistics the final re-ranking of best translation candidates, as will be better explained below. This is how the authors summarize their approach: "a hybrid MT architecture, combining state-of-the-art linguistic processing with advanced stochastic techniques. Grounded in a theoretical reflection on the division of labor between rule-based and probabilistic elements in the MT task … combining component-internal scores and a number of additional sources of (probabilistic) information, … explore discriminative re-ranking of n-best lists of candidate translations through an eclectic combination of knowledge sources";

B. A second way is the one Tan et al. 2012 propose, in their seminal work. They present a new language model which is an "a large scale distributed composite language model that is formed by seamlessly integrating n-gram, structured language model and probabilistic latent semantic analysis under a directed Markov random field paradigm to simultaneously account for local word lexical information, mid-range sentence syntactic structure, and long-span document semantic content". That is, they try to combine semantics/pragmatics, syntax and string-based statistical processing

As for method B., in the abstract to their article they present their approach as follows:

> The composite language model has been trained by performing a convergent N-best list approximate EM algorithm and a follow-up EM algorithm to improve word prediction power on corpora with up to a billion tokens and stored on a supercomputer. The large scale distributed composite language model gives drastic perplexity reduction over n-grams and achieves significantly better translation quality measured by the BLEU score and "readability" of translations when applied to the task of re-ranking the N-best list from a state-of-the-art parsing-based machine translation system. (ibid., 1)

The reason to resort to such an approach reflects the obvious fact that – as the authors note-, the technology based on n-grams has reached a plateau and there is a desperate need to find a new approach to language modeling (Lavie et al. 2006). Work on Chinese has pushed the over of n-gram up to 6-gram obtaining better translations, but the improvement beyond that is minimal (Zhang 2008).

Wang et al. (2006) studied the stochastic properties for a composite language model that integrates n-gram, probabilistic dependency structure in structured language model (SLM), and probabilistic latent semantic analysis (PLSA) under the directed Markov random fields (MRF) framework (Wang et al. 2005). They derived another generalized inside-outside algorithm to train composite n-gram, SLM and PLSA language model from a general EM algorithm by following Jelinek's ingenious definition of the inside and outside probabilities for SLM (Jelinek 2004).

Eventually, the authors are aiming to influence word prediction to find best word pair triggers, with both (dependency) syntactic and (discourse topic) semantic/pragmatic information (Wallach 2006). The output can be combined and used with trigrams in a composite model to drive the final decoder. The resulting language model is defined as the "composite 5-gram/2-SLM+2-gram/4-SLM+5-gram/PLSA1 language model". The interesting part of the evaluation, which as expected is reported

to increase BLEU scores by a 1.19%, is the one dedicated to the "readability" of Chinese-English translations, where they ask human judges to evaluate semantic versus grammatical correctness. In a table the authors report the results of "readability" evaluation on 919 translated sentences of 100 documents, and divide the sentences into four groups: perfect, only semantically correct, only grammatically correct, wrong. The evaluation shows improvements when going from baseline and simple string-based 5-gram processing to the composite language model created by the authors. The greatest relative variation is shown by G(rammatical) sentences, over 60% increase; then P(erfect) sentences, over 50% increase; the lowest relative variation is in S(emantic) sentences that only increase by a 7%. Overall totally wrong sentences decrease by 25%, again a remarkable achievement. This notwithstanding, we can easily see that the amount of "readable" sentences reaches the 66% of the total 919 from a starting point of 56%. However, it is important to stress that they obtain these results by training on a 1.3 billion word corpus using a supercomputer and will not be duplicated on smaller hardware in the near future.

Interesting enough, the results above are almost comparable to those obtained by the LOGON project, in which with a totally different technology and a much smaller corpus, they carried out an evaluation on domain-bounded sentences of unseen running text, they found that on the two thirds (62%) that have been translated, they reached an accuracy of 72.28%. The evaluation carried out by Johannessen et al. 2008 with the help of human judges, based on quality parameters such as "fidelity" and "fluency" showed a result of around 2 points on a graded scale that goes from 0 to 3, where 2 is translated as fair fidelity and still some mistakes in fluency. As the authors themselves comment, the scarsity of resources existing for Norwegian has been the main motivation for building a semantic-transfer-oriented translation system that uses stochastic processing for the target language model, English (see also Llitjós and Vogel 2007). Minimal Recursion Semantics is the "glue" which performs transfer from source to target language and serves as the information vehicle between LFG and HPSG. For a similar approach purely cast in LFG see Riezel and Maxwell 2006. Purely statistical approaches are doomed to failure. In addition the probabilistic NLP experience by itself suggests that a "ceiling" effect has already been reached. As the authors say,

> The Norwegian LOGON initiative capitalizes on linguistic precision for high-quality translation and, accordingly, puts scalable, general-purpose linguistic resources—complemented with advanced stochastic components—at its core. Despite frequent cycles of overly high hopes and subsequent disillusionment, MT in our view is the type of application that may demand knowledge-heavy, 'deep' approaches to NLP for its ultimate, long-term success. (ibid., 144)

Eventually (Bellegarda 2001, 2003) anticipated what is needed, that is a "more polyvalent, multi-faceted, effective and tractable solutions for language modeling – this is only beginning to scratch the surface in developing systems capable of deep understanding of natural language" . In order to achieve this, it is not sufficient to increase the size of data to obtain a breakthrough in the performance. It has been shown that it is not the complicacy of the algorithm that makes the difference: simple algorithms may outperform more complicate ones. However, as Tan et al. have demonstrated, increasing the size of the data has brought improvements, but

then it has been the increase in the complexity of the model that has made the difference. "For language modeling in particular, since the expressive power of simple n-grams is rather limited, it is worthwhile to exploit latent semantic information and syntactic structure that constrain the generation of natural language, this usually involves designing sophisticated algorithms. Of course, this implies that it takes a huge amount of resources to perform the computation." (ibid., 49) Of course, for this to become a feasible alternative, a large scale distributed language model would be required, possibly via cloud computing. Their conclusions are as follows, they intend to

> … construct a family of large scale distributed composite lexical, syntactic, and semantic language models. Finally we'll put this family of composite language models into a phrased-based machine translation decoder that produces a lattice of alternative translations/transcriptions or a syntax-based decoder that produces a forest of alternatives (such integration would, in the exact case, reside in an extremely difficult complexity class, probably PSPACE-complete) to significantly improve the performance of the state-of-the-art machine translation systems.

What really matters at this point is confirming the increasing improvements of SMT while at the same time keeping strictly in mind its inherent limitations. Martin Kay, in one of his latest papers (2011), nicely expresses his pessimism and at the same time optimism towards possible future prospects of MT. He connects MT to what a human translator is doing when translating between English and French, coming to the obvious conclusion that the output is inextricably bound to cultural issues and not just a matter of lexical, semantic or structural knowledge. In fact, by just increasing the size of the training data, one might come up with more cultural problems to solve. Further, the language model would be less adequate the more data are introduced, unless they belong strictly to the same domain, or as he puts it "… new data generally opens at least as many questions as it settles" (ibid., 18). The problem is that "… there is much in any but the most trivial texts that a reader must infer from what is made explicit, but what is explicit in a translation in another language is not generally the same, so that substantive information is both added and subtracted by the translator"(ibid., 15). I will not report his examples, but we can all can try examples on any online available translator to realize the truth of his statement. In fact, what is implicit, is not just "pragmatically" based and culturally motivated, but also in some cases, specifically language related.

We already saw one example above – from English to agreement aware languages like German, French and Italian-, in which researchers discussed questions related to agreement. Now I will turn the other way around. Implicit elements in most cases correspond to "Empty" or "Null" elements as they are usually defined. For instance, in Penn Treebank, they have been manually classified and counted, and the total number for an English treebank is over 36,862 cases of null elements (including traces, expletives, gapping and ambiguity) as listed in Johansson and Nugues (2007), in other words, one every complex sentence, and one every three simple sentences. Then there is the problem of coindexation, or assignment of an antecedent to the empty element: 8,416 are not coindexed, that is 22.83% (see Dienes and Dubey 2003; Schmid 2006). If we exclude all traces of WH and topicalization and limit ourselves to the category OTHER TRACES which includes all unexpressed SBJ of infinitivals

and gerundives, we come up with 12,172 cases of Null non-coindexed elements, 33% of all cases. However, these numbers are this is still a small percentage when compared to languages like Chinese (Cai et al. 2011; Yang and Xue 2010) or some Romance languages like Italian which allow for free null subjects (also objects in Chinese) insertion in tensed clauses. In our treebank of Italian called VIT (Tonelli et al. 2008; Delmonte et al. 2007), we counted an addition of 51.3% of simple sentences with non-canonical, or lexically unexpressed subjects. Obviously this covers the total number utterances in the small corpus (60K tokens) of transcribed spoken dialogues, where the implicit is much higher than in the written text.

Here below are some example translations from Italian to English – but I assume they could easily be from Portuguese, but also from Japanese, and Chinese to English – in which we quite simply demonstrated that the "implicit" (Delmonte 2009a, b), might in many cases determine what is missing in the translation and is, in fact, desperately relevant. Computing complete Predicate-Argument structures is essential for Machine Translation tasks – as Chung and Gildea (2010) have shown where one of the two languages belongs to typology above. As an example, we tried the translation of one sentence from Italian into English, introducing null elements and lexical pronouns, both on Systran and Google online translation websites:

Italian Original

Maria successivamente, dopo aver rifiutato la sua offerta, gli ha detto che vuole vendere la propria casa a sua sorella perché vuole aiutarla.

Gold Translation

Then, after having rejected his offer, Maria told him that she intends to sell her (own) house to her sister because she wants to help her.

Google Translation

Maria later, after she refused his offer, told him he wants to sell his house to his sister because she wants to help.

Systran Translation

Maria successively, after to have refused its offer, she has said it that she wants to sell own house to its sister because she wants to help.

The sentence is fairly simple both in lexical choice and syntactic structure. As one can be gather, Google makes grammatical mistakes due to lack of long distance control – "he, his, his" are all in masculine gender rather than feminine. Systran gets the subject empty pronouns right, but then mistakes the possessives – "its" is neutral – and uses infrequent adverbials like "successively" to translate "dopo". As usual, Google gets an overall best translation both for the grammatical and lexical aspects. Neither of the translation includes the object enclitic "-la"/her in the output. In fact, the verb "help" can be used intransitively, i.e. omitting the object and no mistake ensues. However in this way the leftover pronoun is implicit and needs to be evoked. If we substitute "aiutarla" with "lasciarla" we obtain two different behaviours. Google produces the same output: no pronoun. In this case, however the meaning is no longer preserved and "she wants to leave" has a totally different meaning from

"she wants to leave her". Systran on the contrary produces "it" for singular no matter what gender it is ("lo", "la"), and "them" for plural.

We will take again Kay's preliminary conclusion on the topic to close the section,

> Since examples of the kind just considered are clearly beyond the reach of current, or any readily foreseeable technology — especially if based on machine learning — we must take it that they do nothing but degrade the best performance of the systems that are learned from the texts that contain them. Supervised learning from a corpus of translations that were stricter, if less idiomatic, should surely be expected to result in superior systems. But large corpora of such translations do not occur naturally, would be expensive to produce artificially, and would be required to meet different criteria as the field progressed.

Speech-to-Speech MT

This section fully addresses what, in my opinion, will be the driving application for the future of MT, the one that most users will come to terms with, using mobile devices and other similar technologies. At the heart of Speech-To-Speech MT or S2S for short, there is the need to communicate worldwide orally, for many different kinds of purposes. In other words, we are talking about the need to implement systems for multimodal multilingual communication. This is the future of man–machine interface programs and at the heart of any future development in the associated fields not only of Artificial Intelligence, Speech Synthesis and Automatic Speech Recognition, but also Image Processing, Computer Vision and Face Recognition to be used also in robotics. Thus, MT is only one facet of this important application domain that is based on advancements in basic fields of research like computational linguistics, pattern recognition, and machine learning. Multilingual tools of the future will have to incorporate some if not all of these facilities in order to make real breakthrough in the application market. It is quite obvious to me that a multilingual translation system that is able to take advantage of both spoken and visual input is by far more promising that its companion system that only makes use of written input in a dialogue situation (Zong et al. 2002).

S2S end-to-end systems are organized in a number of complementary modules that receive spoken input in one language and elaborate the corresponding spoken form in another (Karakos and Khudanpur 2008). This is one possible pipeline:

- Speaker produces an utterance in language A, to a device that is linked to an ASR system
- The ASR turns the spoken utterance in its transcribed version still in language A
- A system of utterance and dialogue understanding computes its meaning via an NLP system – this would be the interlingua based approach
- The interlingua is then passed to a Language Generator for language B.
- OR
- The sentence is passed to the MT system that produces the most probable translation into a language B by choosing the best candidate in a list
- The translated sentence is passed to a Speech Synthesizer for language B which speaks it into a device for the user speaker of language B.

Zhang Ying (Joy) (2004) reports in his Survey of Current Speech Translation Research that the translation engines utilized were basically a Multi-Engine MT (MEMT) system,

> … whose primary engines were an Example-Based MT (EBMT) engine and a bilingual dictionary/glossary. Carnegie Mellon's EBMT system uses a "shallower" approach than many other EBMT systems; examples to be used are selected based on string matching and inflectional and other heuristics, with no deep structural analysis. The MEMT architecture uses a trigram language model of the output language to select among competing partial translations produced by several engines. It is used in this system primarily to select among competing (and possibly overlapping) EBMT translation hypotheses. The translated chaplain dialogs provided some of the training. Pre-existing parallel English-Croatian corpora is also used. An addition finite-state word reordering mechanism was added to improve placement of clitics in Croatian. (ibid., 2–3)

Systems like SPEECHLATOR (Waibel et al. 2003), or MASTOR (Gao et al. 2008) work in interlingua modality. All applications have been cast in limited domains and in particular in the hotel room reservation task and have to cope with spontaneous speech. Most importantly, the C-Star consortium, Consortium for Speech Translation Advanced Research, used interlingua, defined as "a computer readable intermediate language that describes the intention of a spoken utterance of a particular language" (ibid., 5), and the domain was related to travel planning. Translating the "intention" allowed system designers to substitute sloppy ramblings in the input or different ways of expressing the same meaning with the one translation available. ***Nespole!*** (Lavie et al. 2002) was one such system which followed another similar system called JANUS III (Levin et al. 2000). Other interesting systems were Digital Olympics (Zong 2008) produced for the Olympics in Peking, and cofunded by the European authorities and the Chinese government, and the NEC Speech Translation System (Yamabana et al. 2003).

The second case of translation is referred to systems like ATROS (Automatic Trainable Recognizer of Speech) developed in the EuTrans project, which aim to synchronize speech recognition models with linguistic levels like lexical, syntactic and eventually translation model, by the use of FST. The MT technique followed is example-based. The AT&T approach uses multimodal parsing and understanding always with a finite-state model. The system subdivides the translation task into a phase for lexical choice and another phase for lexical reordering (ibid., 5). The lexical choice phase is divided up into phrase-level and sentence-level using different translation models. Eventually, the reordering phase approximates a tree-transducer using a string transducer.

To describe current state-of-the-art STSMT we will be referring to the international challenge and associated evaluation campaign called QUAERO, reported in Lamel et al. 2011, for a bidirectional French-German task, which has seen the participation of the most important actors on the scene: RWTH, KIT, LIMSI and SYSTRAN. In the Reference section websites of the partners involved in the common task are reported.

Problems related to STSMT are common and different from standard written-based SMT. First of all, there is the need to make available a recognition vocabulary

big enough to include all word-forms possibly present in the task at hand in order to reduce the number of Out Of Vocabulary (OOV) rates (see Gales et al. 2007). This constitutes the worst problem to solve, and it is not just a matter of increasing a list, with frequency of occurrence. Words included in the recognition vocabulary needs to be represented phonetically. Vocabulary sizes range from 65K to 300K words as reported in Lamel et al. and OOV rates range from around 0.5% to 2%. It is interesting to note that systems represent the pronunciation dictionary with sets of phone symbols that go from 35 up to a maximum of 50 symbols. Systems generate the phonetic representation with different methods: some use rule based grapheme to phoneme conversion, others statistical methods, or a combination of the two, often introducing a list of manually verified exceptions. Most phone sets include pseudo phones for silence and non-speech sounds and there are typically 1.1–1.3 pronunciations per word. However, as will be explained below, there are special provisions for prosodically related pronunciation variants, which in a language like French, constitute a frequent phenomenon with which to cope.

In STSMT in addition to language models and translation models, there are acoustic models to build. In particular, current acoustic models are trained on several hundreds of hours of audio data coming from a variety of sources. Language models are trained on over a billion words of texts, comprised of assorted newspaper and newswire texts: they end up typically containing around 400M 4-grams for both languages.

It is important to note that all speech recognition experiments described in Lamel et al. were performed with the help of the Janus Recognition Toolkit (JRTk) from CMU and the Ibis single pass decoder described in Soltau et al. 2001. The Janus system is described in (Lavie and Waibel 1996; Levin et al. 2000). Training for German was performed using some 350 h of training material from different sources. As reported in Lamel et al.

> Two different front-ends were applied: The warped minimum variance distortionless response (WMVDR) approach and the conventional (Mel-frequency Cepstral Coefficients) MFCC approach. The front-end uses a 42-dimensional feature space with linear discriminant analysis and a global semi-tied covariance (STC) transform with utterance-based cepstral mean and variance normalization. The 42-dimensional feature space is based on 20 cepstral coefficients for the MVDR system and on 13 cepstral coefficients for the MFCC system … All the acoustic data is in 16 kHz, 16 bit quality. Acoustic model training was performed with fixed state alignments and Vocal Tract Length Normalization (VTLN) factors … The system uses left-to-right hidden Markov Models (HMM)s without state skipping with three HMM states per phoneme.

This produced an adapted gender- and speaker-independent acoustic model. The Language Model for German was built from a variety of text sources and resulted in a 10GB LM, containing 31.7M 2-grams, 91.9M 3-grams, 160.4M 4-grams, as reported in the same paper. Speaker adaptation was performed in a second pass and produced FSA-SAT models with language models which were even bigger.

For French, approximately 330 h of audio data were used to train the acoustic models. The segmentation was implemented in two steps applying an HMM-based segmenter which took into consideration different speech events, noises, silences and music. For each speech segment a Gaussian Mixture Model is generated. Two different

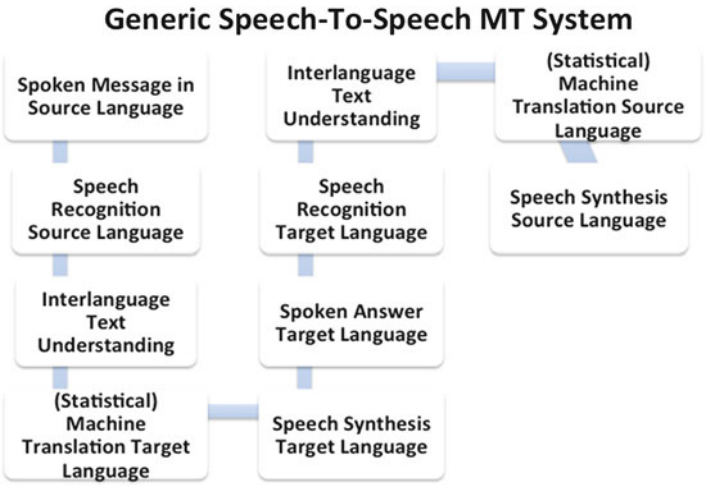

Fig. 6.3 Generic pipeline for a STSMT system

kinds of phoneme sets are used for training: a first one consisting of 35 phonemes and another version provided by Vocapia that consists of 32 phonemes. As with German. two types of acoustic front-end were used: one based on Mel-frequency Cepstral Coefficients and the other one on the warped minimum variance distortionless response. Both front-ends work on a window of 10ms. At the end of the training process five acoustic models were produced, which were improved by boosted Maximum Mutual Information Estimation training. LMs were trained with all tests available using the SRI Language Modeling Toolkit. Interesting enough, for the training procedure a dictionary was used which contained hesitations, fragments, human and non-human noise in addition to pronunciation variants of each word, taken from GlobalPhone and Lexique 3. Missing pronunciations were generated with Sequitur G2P (Bisani and Ney 2008). The final system reached 27% WER on the Quaero Development Set. As reported in the Conclusions, there has been a steady progress in reducing the WER in the last 3 years. For some languages reduction reaches 25%, and it is around 15% for the three primary languages – French, German English (Fig. 6.3).

TED Conferences Talks and the WIT3 Corpus

TED Conference at TED (Technology, Entertainment and Design) website, www.ted.com, have been posting video recording of talks, having as their subject cultural issues in general. Talks come with English subtitles and their translations in more than 80 languages. This has been done since 2007 for the sake of sharing ideas around the world, as the organizer comment on the website.

FBK (Bruno Kessler Foundation) in Trento (Italy) have organized a website https://wit3.fbk.eu/, with the aim of redistributing the whole corpus with original textual

contents in their multilingual transcriptions, but also to make a ready-to-use version with MT benchmarks and processing tools for research purposes. The acronym of the website stands for Web Inventory of Transcribed and Translated Talks. A detailed description of the corpus can be found in a recent paper by Cettolo et al. 2012.

The same organizers, including Marcello Federico, Mauro Cettolo together with Michael Paul from NICT (Japan) and Sebastian Stueker (KIT) Germany, are responsible for the TED Task Evaluation Campaign which is an important event related to IWSLT conferences, and can be found at http://iwslt2012.org/, subdirectory "evaluation-campaign/ted-task". TED Task includes the following subtasks:

> IWSLT proposes challenging research tasks and an open experimental infrastructure for the scientific community working on spoken and written language translation. The IWSLT 2012 Evaluation Campaign includes the TED Task, that is the translation of TED Talks, a collection of public speeches on a variety of topics. Three tracks are proposed addressing different research tasks:
>
> ASR track : automatic transcription of talks from audio to text (in English)
>
> SLT track: speech translation of talks from audio (or ASR output) to text (from English to French)
>
> MT track : text translation of talks for two language pairs plus eight optional language pairs:
>
> official: from English to French and from Arabic to English
>
> optional: from German, Dutch, Polish, Portuguese-Brazil, Romanian, Russian, Turkish and Chinese to English
>
> Main challenges of the proposed tracks are:
>
> Open domain ASR, clean transcription of spontaneous speech, detection and removal of non-words, and talk style and topic adaptation.
>
> Open domain SLT, translation of speech or ASR output into true-case punctuated text, and talk style and topic adaptation.
>
> Open domain MT between distant languages, and talk style and topic adaptation.

Full guidelines can be found on the same website. This is certainly not the only initiative – KIT, Germany would be another one – but certainly one of the most important ones.

The GALE DARPA MT Project

As discussed at the beginning of this chapter, a number of different institutions are currently contributing to finance research efforts of the vast community of scientists working in the field of MT. The most important of these international initiatives in favour of the improvement of MT research is the one financed by DARPA. The project was originally called GALE and has recently partially concentrated on BOLT (Broad Operational Language Translation), which has clear military goals. As to this project, the interesting thing that happened last year, was the hiring of Professor Bonnie Dorr as manager of BOLT: this event has an extremely important meaning. It testifies to the switch of perspective the DARPA management wants to give to the project: from statistics only, to the massive introduction of linguistics, that is syntax and semantics, into MT. Of course we regard this change of point of view a successful move towards finding the best approach for the MT of the future.

To comment on the GALE (Global Autonomous Language Exploitation) program we will be referring basically to the original webpage dedicated to the project directly on the DARPA website and from a presentation downloadable from their general internal search engine. DARPA has been a key sponsor of machine translation, as well as computer processing of language (speech and text) work for over three decades. GALE began in September 2005 and is still continuing even though it was scheduled to run only until September 2010. GALE speech-to-text and machine-translation researchers came from the following corporations and organizations, as reported by The Associated Press, in a 2006 online article on the topic: IBM Corp., backed by a $6 billion annual research budget; SRI International, a $300 million, nonprofit research organization based in Silicon Valley; and BBN Technologies Inc., a $200 million research contractor headquartered in Cambridge BBN nabbed people at Cambridge University, the universities of Maryland and Southern California and a French lab, among others. IBM got researchers from Carnegie Mellon, Johns Hopkins, Brown University and Stanford, plus other researchers at the University of Maryland. SRI's links included European and Asian schools, Columbia University and the universities of California and Washington. The goal is to create technology that will automatically translate spoken or written words from foreign languages into grammatically correct English. GALE has an ambitious goal of reaching 95% accuracy without human mediation. The technology is moving toward allowing the translations to happen in real time. These goals are set forth in ambitious research efforts and according to DARPA "these efforts are poised to come close to achieving their goals in certain specific contexts with Modern Standard Arabic and Mandarin Chinese". The largest of these efforts was the 5-year, multimillion-dollar-per-year GALE program, which seeks real-time translation of Modern Standard Arabic and Chinese print, Web, news, and television feeds. The second program is the Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program. TRANSTAC's goal is highly specific: a portable two-way speech translation system that enables an average soldier to communicate with a person who cannot speak English. NIST Machine Translation Evaluation for GALE: The Speech Group in the Information Technology Laboratory's Information Access Division at NIST is undertaking the development of an evaluation for machine translation (MT) using "edit-distance" as the evaluation metric as defined by the GALE program. The GALE program will evaluate MT in terms of the quality of the system translations. This will be accomplished by measuring the edit distance between a system output and a gold standard reference. The term "edit-distance" refers to the number of edits (modifications) that someone (human) needs to make to the output of a machine translation system such that the resulting text is fluent English and completely captures the meaning of the gold standard reference. In order to achieve its goal, the GALE program will have to develop and apply computer software technologies to analyze and interpret huge volumes of speech and text. Output information will have to be in easy-to-understand forms for military personnel and monolingual English-speaking analysts to use, in response to direct or implicit requests. GALE will consist of three major engines: Transcription, Translation and Distillation. The output of each engine will be limited

to English text. The distillation engine integrates information of interest to its user from multiple sources and documents. Military personnel will interact with the distillation engine via interfaces that could include various forms of human-machine dialogue (not necessarily in natural language).

According to the project description on the LDC website, the Linguistic Data Consortium supports the GALE Program by providing linguistic resources – data, annotations, tools, standards and best practices limited to Arabic – for system training, development and evaluation. The Translation process will take Arabic source text drawn from many different genres, both spoken and written, and translate it (hopefully) into fluent English while preserving all of the meaning present in the original Arabic text. Translation agencies will use their own best practices to produce high quality translations, according to specific guidelines so that all translations are guided by some common principles. Linguistic Data Consortium will also be providing post-editing of MT output in order to compute "edit distance" between machine translations and human gold standard translations. The post editor's role is to compare computer-translated texts against the same texts translated by humans. Working with one sentence at a time, the editor modifies the computer translation until its meaning is identical to the human translated sentence.

As indicated above, GALE's goal was to deliver, by 2010, software that can almost instantly translate Arabic and Mandarin Chinese with 90–95% accuracy. Fortunately for the GALE teams, they didn't have to be near 95% right away. In the first year, they were expected to translate Arabic and Mandarin speech with 65% accuracy; with text the goal was 75%. In an interview reported on SLATE, Mari Maeda, a DARPA manager who ran the program, says that, by the end, "TransTac achieved about 80% accuracy: enough to be interesting, but not enough to be useful." Considering state of the art MT field that was already to be regarded as a particularly difficult goal to achieve. DARPA estimated that the best systems could translate foreign news stories at 55% accuracy. But DARPA wanted translations not only from such controlled, well-articulated sources: in fact, it was to be considered as an open domain, unlimited vocabulary task. As reported in The Associated Press article "GALE incorporates man-on-the-street interviews and raucous colloquial chats on the Web. Background noise, dialects, accents, slang, short words … that most speakers don't bother to clearly enunciate – these are the stuff of nightmares for speech-recognition and machine-translation engineers". We have presented and discussed at length all of these "nightmares" in the chapter. The test – hours of audio and dozens of documents in Arabic and Mandarin – and the evaluation was done by counting the number of human edits that the sentences needed in order for them to have the correct meaning. Again, quoting from the TAP article, "the results largely met DARPA's demands of 75% accuracy for text translation and 65% for speech… The BBN-led team produced 75.3% accuracy with Arabic text, 75.2% in Chinese. It scored 69.4% in Arabic speech; 67.1% in Mandarin. IBM scored higher with Arabic text and SRI scored higher in Mandarin. The current successor of TransTac is called BOLT and Bonnie Dorr, program manager forBOLT, says that DARPA is now "very focused on moving beyond statistical models." The reason is that, as you throw more and more paral-

lel data at your algorithms, you "get diminishing returns. The payoff gets smaller, and you start to plateau with your results even if you increase the volume of training data." (again reported on SLATE online).

If DARPA is financing LDC to produce additional translation of Arabic, this is a clear sign of two symptoms:

- More training data are required
- The current results of translation systems are still unsatisfactory
- Of course, we assume that a final strategic improvement will not come until another additional piece of research is added to the list:
- Inventing new translation and language models that will incorporate semantics and pragmatics (besides syntax) in a most fruitful way, which is not yet the case
- Combining more systems in a pipeline to produce a hybrid hypersystem that can learn

This is clearly what the MT community as a whole is striving for and what I assume will happen in future.

Conclusion

In this chapter I have presented and commented what I regard the most interesting technologies and methodologies for Machine Translation, including both Rule-Based and Statistically-Based systems. I devoted a first section to purely statistically based systems, highlighting their shortcomings and their advantages which make them more and more important for the future of MT. In fact, a statistical model is shown to be an essential component of most Rule-Based systems reviewed in another section: these systems take advantage of the ability of generalization that statistics makes easily available to create what are usually called hybrid systems. A third section has been devoted to syntactically based systems which make use of syntactic tree structures of dependency structures to produce better modelling of the data and hopefully better translations. These systems are strongly dependent on computational linguistic tools like parsers, morphological analysers and suffer from their shortcomings which impinge on the final translation accuracy level.

I take graph-based models to be superior in general to phrase-based or word-based models, the reason being simply the fact that structural properties of both input and output can be duly taken into consideration in the modelling statistical phase. Why this is important should now be clear: in order to produce a real step forward, Machine Translation should incorporate properties belonging to both syntax and semantics of the sentence and text to be translated. This can only be achieved with a structurally aware statistical model. I take models relying on dependency structure to be the reference point with additional constraints however: the need to satisfy predicate-argument restrictions as realized in the statistical graph-based model and reinforced in the observed data.

Rule-Based systems could still be useful to encode Multiwords correspondences and other idiomatic expressions which require some preprocessing. However, as the case of feature agreement has clearly shown, Rules-Based systems are unable to enforce local matching requirements when fine-grained coupling is needed. They could perhaps work in postprocessing to check such local agreements with highly targeted rule systems, which must be language dependent.

Eventually, speech-to-speech multilingual processing has started to be used in real life applications. This is great news, but also bad news as far as current achievements are concerned. The effort however is enormous and the quantity of resources in play is outside the scope of any single researcher computing ability. Only specialized centers may legitimately aim at competing in such international challenges. What about the use of visual computing and the interpretation of facial movements or other additional gestures? Multimodal computation is still in its infancy and its interaction with natural language processing tools is expected to grow in the future. As to current situation, I don't know of any system capable of taking advantage of gestures of facial expressions to improve its multilingual tools.

New mathematical models are needed that incorporate all types of knowledge needed to come as close as possible to what human translators do: best translators are always domain constrained, and this applies to both humans and computers. Syntax poses different challenges from pragmatics and semantics: new mathematical models need to take these differences into adequate account.

Last but not least, the need to foster improvements in the companion field of computational linguistics, which alone can come up with complete linguistic representations needed in the Rule-Based scenario. I am referring to the need to enrich dependency structures with null elements and annotate them with coreference information to allow for proper agreement features to be instantiated. Such new tools could then be used to produce Logical Form representations to better handle meaning differences and ambiguities.

References

Abney S (1989) Parsing by chunks. In: Tenny C (ed) The MIT parsing volume, 1988–89. Center for Cognitive Science, MIT, Cambridge

Adly N, Alansary S (2009) Evaluation of Arabic machine translation system based on the universal networking language. In: The 14th international conference on applications of natural language to information systems "NLDB 2009", Saarland University, Saarbrucken, Germany, 23–26 June 2009

Alansary S, Nagi M, Adly N (2009) The universal networking language in action in English-Arabic machine translation. In: 9th conference on language engineering, Cairo, pp 1–12

Alegria I, Diaz de Ilarraza A, Labaka G, Lersundi M, Mayor A, Sarasola K (2007) Transfer-based MT from Spanish into Basque: reusability, standardization and open source, vol 4394, Lecture Notes in Computer Science. Springer, Berlin/New York, pp 374–384

Alexandrescu A, Kirchhoff K (2009) Graph-based learning for statistical machine translation, 2009. In: Human language technologies: the 2009 annual conference of the North American Chapter of the ACL, Boulder, Colorado, pp 119–127

Alkuhlani S, Habash N (2011) A corpus for modeling morpho-syntactic agreement in Arabic: gender, number and rationality. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics (ACL'11), Portland, Oregon, USA

Ambati V, Lavie A (2008) Improving syntax driven translation models by restructuring divergent and non-isomorphic parse tree structures. In: Proceedings of the eighth conference of the association for machine translation in the Americas, Waikiki, pp 235–244

Ambati V, Lavie A, Carbonell J (2009) Extraction of syntactic translation models from parallel data using syntax from source and target languages. In: MT Summit XII: proceedings of the twelfth machine translation summit, Ottawa, 26–30 Aug 2009, pp 190–197

Ambati V, Vogel S, Carbonell J (2011) Multi-strategy approaches to active learning for statistical machine translation. Associated press article on the web: http://www.cnn.com/2006/TECH/11/06/darpa.translation.ap/index.html

Apidianaki M (2009) Data-driven semantic analysis for multilingual WSD and lexical selection in translation. In: Proceedings of the 12th conference of the European chapter of the association for computational linguistics (EACL), Athens, 30 Mar–3 Apr 2009, pp 77–85

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

Ayan NF, Zheng J, Wang W (2008) Improving alignments for better confusion networks for combining machine translation systems. In: Proceedings of the Coling'08, pp 33–40

Baker K, Bloodgood M, Dorr BJ, Callison-Burch, C, Filardo, NW, Piatko, C, Levin L, Miller S (2011) Modality and negation in SIMT – use of modality and negation in semantically-informed syntactic MT. Comput Linguist 38(2):1–48, (accepted for publication)

Baker K, Bloodgood M, Dorr BT, Callison-Burch C, Filardo NW, Piatko C, Levin L, Miller S (2012) Modality and negation in SIMT – use of modality and negation in semantically-informed syntactic MT. Comput Linguist 1:1–48

Banchs, RE, Costa-jussà MR (2010) A non-linear semantic mapping technique for cross-language sentence matching. In: Proceedings of the 7th international conference on advances in natural language processing (IceTAL), Reykjavik, pp 57–66

Banchs RE, Costa-jussà MR (2011) A semantic feature for statistical machine translation. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, ACL HLT 2011, Portland, June 2011, pp 126–134

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Ann Arbor, June, pp 65–72

Banerjee P, Dandapat S, Forcaday ML, Groves D, Penkale S, Tinsley J, Way A (2011) Technical report: OpenMaTrEx, a free, open-source hybrid data-driven machine translation system

Béchara, H, Ma Y, van Genabith J (2011) Statistical post-editing for a statistical MT system. MT Summit XIII: the thirteenth machine translation summit [organized by the] Asia-Pacific association for machine translation (AAMT), Xiamen, 19–23 Sept 2011, pp 308–315

Bellegarda J (2001) Robustness in statistical language modeling: review and perspectives. In: Junqua J, van Noods G (eds) Robustness in language and speech technology. Kluwer, Dordrecht/Boston, pp 101–121

Bellegarda J (2003) Statistical language model adaptation: review and perspectives. Speech Commun 42:93–108

Bisani M, Ney H (2008) Joint sequence models for grapheme-to-phoneme conversion. Speech Commun 50(5):434–451

Blackwood G, de Gispert A, Byrne W (2008) Phrasal segmentation models for statistical machine translation. In: Proceedings of 22nd international conference on computational linguistics (COLING), Manchester

Blain F, Senellart J, Schwenk H, Plitt M, Roturier J (2011) Qualitative analysis of post-editing for high quality machine translation. In: Proceedings of MTS: 13th machine translation summit, Xiamen, pp 164–171

Bontcheva K (2005) Generating tailored textual summaries from ontologies. In: The semantic web: research and applications, Springer, pp 531–545

Brown RD (1996) Example-based machine translation in the Pangloss system. In: Proceedings of the 16th international conference on computational linguistics (COLING-96), Copenhagen, pp 169–174

Brown PF, Della Pietra SA, Della Pietra VJ, Goldsmith MJ, Hajic J, Mercer RL, Mohanty S (1993) But dictionaries are data too. In: Human language technology: proceedings of a workshop held at Plainsboro, New Jersey, USA, Morgan Kaufmann, San Francisco, 21–24 March 1993, pp 202–205

Cai S, Chiang D, Goldberg Y (2011) Language-independent parsing with empty elements. In: Proceedings of the 49th annual meeting of the ACL, Portland, pp 212–216

Callison-Burch C, Koehn P, Osborne M (2006) Improved statistical machine translation using paraphrases. In: HLT-NAACL 2006: proceedings of the human language technology conference of the North American chapter of the ACL, New York, June 2006, pp 17–24

Carl M, Way A (eds) (2003) Recent advances in example-based machine translation. Kluwer, Dordrecht. Introduction to the workshop on EBMT, xxxi

Carpuat M, Diab M (2010) Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In: Proceedings of HLT-NAACL 2010, Los Angeles, pp 242–245

Carpuat M, Wu D (2007) Improving statistical machine translation using word sense disambiguation. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007), Prague, pp 61–72

Caseli HM, Nunes MGV, Forcada ML (2006) Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. Mach Trans 20(4):227–245

Cettolo M, Bertoldi N, Federico M (2011) Methods for smoothing the optimizer instability in SMT. In: Proceedings of the 13th machine translation summit, Asia-Pacific Association for Machine Translation, pp 32–39

Cettolo M, Girardi C, Federico M (2012) WIT3: web inventory of transcribed and translated talks. In: Proceedings of EAMT, Trento, Italy, pp 261–268

Chang Pi-Chuan, Huihsin Tseng, Jurafsky D, Manning CD (2011) Discriminative reordering with Chinese grammatical relations features. In: Proceedings of SSST-3, third workshop on syntax and structure in statistical translation, pp 51–59

Chao Wang, Collins M, Koehn P (2007) Chinese syntactic reordering for statistical machine translation. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), Prague, Czech Republic, pp 737–745

Chappelier J-C, Rajman M (1998) A generalized CYK algorithm for parsing stochastic CFG. In: Proceedings of tabulation in parsing and deduction (TAPD'98), Paris, France

Chen Stanley F, Goodman JT (1998) An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University

Chen Yu, Eisele A (2010) Integrating a rule-based with a hierarchical translation system. In: LREC 2010: proceedings of the seventh international conference on language resources and evaluation, Valletta, Malta, 17–23 May 2010, pp 1746–1752

Chenqing Zong, Bo Xu, Taiyi Huang (2002) Interactive Chinese-to-English speech translation based on dialogue management. In: Proceedings of the workshop on speech-to-speech translation: algorithms and systems, pp 61–68

Cherry C, Lin D (2003) A probability model to improve word alignment. ACL-2003: 41st annual meeting of the Association for Computational Linguistics, Sapporo, Japan, 7–12 July 2003

Chiang D (2005) A hierarchical phrase-based model for statistical machine translation. In: Proceedings of ACL, pp 263–270 (Best paper award)

Chiang D (2007) Hierarchical phrase-based translation. Comput Linguist 33(2):202–228

Chiang D (2010) Learning to translate with source and target syntax. In: Proceedings of ACL10, Stroudsburg, PA, USA, pp 1443–1452

Chiang D, Marton Y, Resnik P (2008) Online large-margin training of syntactic and structural translation features. In: EMNLP 2008: proceedings of the 2008 conference on empirical methods in natural language processing, Honolulu, 25–27 Oct 2008, pp 224–233

Chiang D, Knight K, Wang W (2009) 11,001 new features for statistical machine translation. In: Proceedings of HLT-NAACL09, Boulder, pp 218–226
Choi JD, Palmer M, Nianwen Xue (2009) Using parallel propbanks to enhance word-alignments. In: Proceedings of ACL-IJCNLP workshop on linguistic annotation (LAW'09), pp 121–124
Chung Tagyoung, Gildea D (2010) Effects of empty categories on machine translation. In: Proceedings of EMNLP, pp 636–645
Cicekli I, Altay Güveniri H (2001) Learning translation templates from bilingual translation examples. Appl Intell 15(1):57–76
Cimiano P, Montiel-Ponsoda E, Buitelaar P, Espinoza M, Gomez-Pérez A (2010) A note on ontology localization. J Appl Ontology 5:127–137
Clark J, Dyer C, Lavie A, Smith N (2011) Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In: Proceedings of ACL, Portland
Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo Q-H, Dien D, Kawtrakul A, Takeuchi K, Shigematsu M, Taniguchi K (2008) Bio-Caster: detecting public health rumors with a web-based text mining system. Bioinformatics 24(24):2940–2941
Cowan B, Kučerová I, Collins M (2006) A discriminative model for tree-to-tree translation. In: EMNLP-2006: proceedings of the 2006 conference on empirical methods in natural language processing, Sydney, July 2006, pp 232–241
Cui Lei, Dongdong Zhang, Mu Li, Ming Zhou, Tiejun Zhao (2010) A joint rule selection model for hierarchical phrase-based translation. In: ACL 2010: the 48th annual meeting of the association for computational linguistics, Proceedings of the conference short papers, Uppsala, 11–16 July 2010, pp 6–11
Cui Lei, Dongdong Zhang, Mu Li, Ming Zhou (2011) Function word generation in statistical machine translation systems. MTS:139–146
Daelemans W, Hoste V (eds) (2009) Evaluation of translation technology. Artesis University College, Department of Translators & Interpreters, Antwerp, 261 p
Dan Melamed I (2004) Statistical machine translation by parsing. In: Proceedings of the ACL 2004: 42nd annual meeting of the association for computational linguistics, Barcelona, 21–26 July 2004, pp 653–660
Dandapat S, Forcada ML, Groves D, Penkale S, Tinsley J, Way A (2010) OpenMaTrEx: a free/open-source marker-driven example-based machine translation system. In: Loftsson H et al (eds) Advances in natural language processing: 7th international conference on NLP, IceTAL 2010, Reykjavík, 16–18 Aug 2010. College lecture notes in artificial intelligence, vol 6233. Springer, Berlin/Heidelberg, pp 121–126
Dandapat S, Morrissey S, Way A, Forcada ML (2011) Using example-based MT to support statistical MT when translating homogeneous data in a resource-poor setting, 2011. In: Forcada ML, Depraetere HS, Vandeghinste V (eds) Proceedings of the 15th conference of the European association for machine translation, pp 201–208
DARPA project commented at: 1. http://www.slate.com/articles/technology/future_tense/2012/05/darpa_s_transtac_bolt_and_other_machine_translation_programs_search_for_meaning_.html. 2. http://www.darpa.mil/Our_Work/I2O/Personnel/Dr_Bonnie_Dorr.aspx. 3. http://www.darpa.mil/NewsEvents/Releases/2011/2011/04/19_DARPA_initiates_overarching_language_translation_research_Publishes_Broad_Agency_Announcement_for_Broad_Operational_Language_Translation_program.aspx
David C (2007) Hierarchical phrase-based translation. Comput Linguist 33(2):202–228
David C (2010) Learning to translate with source and target syntax. In: Proceedings of ACL10, Morristown, pp 1443–1452
de Marneffe M-C, Manning CD (2008) Stanford typed hierarchies representation. In: Proceedings of the COLING workshop on cross-framework and cross-domain parser evaluation
de Marneffe, M-C, MacCartney B, Manning CD (2006) Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC-06
Declerck T, Krieger H-U, Thomas SM, Buitelaar P, O'Riain S, Wunner T, Maguet G, McCrae J, Spohr D, Montiel-Ponsoda E (2010) Ontology-based multilingual access to financial reports for sharing business knowledge across Europe. In: Rooz J, Ivanyos J (eds) Internal financial

control assessment applying multilingual ontology framework, HVG Press Kft., Budapest, pp 67–76
Delmonte R (2009b) A computational approach to implicit entities and events in text and discourse. In: International journal of speech technology (IJST), Springer, pp 1–14
Delmonte R (2009a) Understanding implicit entities and events with Getaruns. In: ICSC, 2009 IEEE international conference on semantic computing, Berkeley, pp 25–32
Delmonte R, Bristot A, Tonelli S (2007) VIT – Venice Italian Treebank: syntactic and quantitative features. In: De Smedt K, Hajic J, Kübler S (eds) Proceedings of sixth international workshop on Treebanks and linguistic theories, Nealt proceedings series vol 1, ISSN 1736–6305, pp 43–54
DeNeefe S, Knight K, Wang W, Marcu D (2007) What can syntax-based MT learn from phrase-based MT? In: Proceedings of EMNLP-CoNLL, pp 755–763
Denkowski M, Lavie A (2010) METEOR-NEXT and the METEOR paraphrase tables: improved evaluation support for five target languages. In: ACL 2010: joint fifth workshop on statistical machine translation and MetricsMATR. Proceedings of the workshop, Uppsala University, Uppsala, 15–16 July 2010, pp 339–342
Deyi Xiong, Min Zhang, Haizhou Li (2010) Learning translation boundaries for phrase-based decoding. In: Proceedings of HLT-NAACL 2010
Dienes P, Dubey A (2003) Antecedent recovery: experiments with a trace tagger. In: Proceedings of EMNLP, Sapporo
Ding Yuan, Palmer M (2005) Machine translation using probabilistic synchronous dependency insertion grammars. In: ACL-2005: 43rd annual meeting of the association for computational linguistics, University of Michigan, Ann Arbor, 25–30 June 2005, pp 541–548
Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Marcus M (ed) HLT 2002: human language technology conference: proceedings of the second international conference on human language technology research, San Diego, 24–27 Mar 2002, [Morgan Kaufmann for DARPA, San Francisco], pp 138–145
Dugast L, Senellart J, Koehn P (2007) Statistical post-editing on SYSTRAN's rule-based translation system. In: Proceedings of the second workshop on statistical machine translation, Prague, pp 220–223
Dugast L, Senellart J, Koehn P (2008) Can we relearn an RBMT system? In: Proceedings of the third workshop on statistical machine translation, Columbus, pp 175–178
Dugast L, Senellart J, Koehn P (2009) Selective addition of corpus-extracted phrasal lexical rules to a rule-based machine translation system. In: MT Summit XII: proceedings of the twelfth machine translation summit, Ottawa, 26–30 Aug 2009, pp 222–229
Dyer C, Muresan S, Resnik P (2008) Generalizing word lattice translation. In: Proceedings of ACL, Columbus, pp 1012–1020
Ebling S, Way A, Volk M, Naskar SK (2011) Combining semantic and syntactic generalization in example-based machine translation. In: Forcada ML, Depraetere H, Vandeghinste V (eds) Proceedings of the 15th conference of the European association for machine translation, Leuven, pp 209–216
Eisele A, Federmann C, Uszkoreit H, Saint-Amand H, Kay M, Jellinghaus M, Hunsicker S, Herrmann T, Chen Y (2008) Hybrid machine translation architectures within and beyond the Euro Matrix project. In: Hutchins J, von Hahn W (eds) Proceedings of EAMT 2008: 12th annual conference of the European association for machine translation, Hamburg, 22–23 Sept 2008, pp 27–34
Eisner J (2003) Learning non-isomorphic tree mappings for machine translation. ACL-2003: 41st annual meeting of the association for computational linguistics, Sapporo, 7–12 July 2003
El Kholy A, Habash N (2010) Orthographic and morphological processing for English-Arabic statistical machine translation. In: Proceedings of Traitement Automatique du Langage Naturel (TALN-10). Montréal, Canada
España-Bonet C, Jesús G, Lluís M (2009) Discriminative phrase-based models for Arabic machine translation. ACM Trans Asian Lang Info Process J 8(4):1–20
España-Bonet C, Labaka G, Dìaz de Ilarraza A, Màrquez L (2011) Hybrid machine translation guided by a rule-based system, MTS, pp 554–561

Espinoza M, Gomez-Pérez A, Mena E (2008) Enriching an ontology with multilingual information. In: Proceedings of the 5th annual of the European semantic web conference (ESWC08), Tenerife, pp 333–347

Espinoza M, Montiel-Ponsoda E, Gomez-Pérez A (2009). Ontology localization. In: Proceedings of the 5th international conference on knowledge capture (KCAP09), pp 33–40

Esplà-Gomis M, Saànchez-Cartagena VM, Pérez-Ortiz JA (2011) Multimodal building of monolingual dictionaries for machine translation by non-expert users, In: Proceedings of the 13th MTS, Xiamen, pp 147–154

Fai Wong, Dong-Cheng Hu, Yu-Hang Mao, Ming-Chui Dong, Yi-Ping Li (2005) Machine translation based on constraint-based synchronous grammar. In: Proceedings of the 2nd international joint conference on natural language, Jeju Island, Republic of Korea, pp 612–623

Federmann C, Eisele A, Uszkoreit H, Chen Y, Hunsicker S, Xu J (2010) Further experiments with shallow hybrid MT systems. In: ACL 2010: joint fifth workshop on statistical machine translation and MetricsMATR. Proceedings of the workshop, Uppsala University, Uppsala, 15–16 July 2010, pp 77–81

Feifei Zhai, Jiajun Zhang, Yu Zhou, Chengqing Zong (2011) Simple but effective approaches to improving tree-to-tree model, MTS

Font-Llitjòs A, Carbonell JG, Lavie A (2005) A framework for interactive and automatic refinement of transfer-based machine translation. In: European association of machine translation (EAMT) 10th annual conference, Budapest, Hungary, Citeseer

Forcada ML, Depraetere H, Vandeghinste V (eds) (2011) Proceedings of the 15th conference of the European association for machine translation, Leuven, pp 13–20

Forcada ML, Ginest ı-Rosell M, Nordfalk J, O'Regan J, Ortiz-Rojas S, Pérez-Ortiz JA, Sànchez-Martìnez F, Ramìrez-Sànchez G, Tyers FM (2011a) Apertium: a free/open-source platform for rule-based machine translation. Machine translation. Special issue on free/open-source machine translation (in press)

Forcada ML, Ginest **á**-Rosell M, Nordfalk J, O'Regan J, Ortiz-Rojas S, Pérez-Ortiz JA, Sànchez-Martìnez F, Ramìrez-Sànchez G, Tyers FM (2011b) Apertium: a free/open-source platform for rule-based machine translation. Mach Translat (Special Issue on freeo/open-source machine translation) 25(2):127–144.

Fraser A, Marcu D (2006) Semi-supervised training for statistical word alignment. In: Proceedings of ACL, Sydney, pp 769–776

Fraser A, Marcu D (2007) Getting the structure right for word alignment: LEAF. In: Proceedings of EMNLP, Prague, pp 51–60

Fraser A, Marcu D (2007b) Measuring word alignment quality for Statistical Machine Translation. Comput Linguist 33(3):293–303

Fu B, Brennan R, O'Sullivan D (2010) Cross-lingual ontology mapping and its use on the multilingual semantic web. In: Proceedings of the 1st workshop on the multilingual semantic web, at the 19th international World Wide Web Conference (WWW 2010)

Gales MJF, Liu X, Sinha R, Woodland PC, Yu K, Matsoukas S, Ng T, Nguyen K, Nguyen L, Gauvain J-L, Lamel L, Messaoudi A (2007) Speech recognition system combination for machine rranslation. In: IEEE international conference on acoustics, speech and signal processing, Honolulu, pp 1277–1280

Galley M, Hopkins M, Knight K, Marcu D (2004) What's in a translation rule? In: HLT-NAACL 2004: human language technology conference and North American chapter of the association for computational linguistics annual meeting, The Park Plaza Hotel, Boston, 2–7 May 2004, pp 273–280

Galley M, Graehl J, Knight K, Marcu D, DeNeefe S, Wang Wei, Thayer I (2006) Scalable inference and training of context-rich syntactic translation models. In: Proceedings of the international conference on computational linguistics/association for computational linguistics (COLING/ACL-06), Sydney, pp 961–968

Gao Q, Vogel S (2011) Utilizing target-side semantic role labels to assist hierarchical phrase-based machine translation. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, ACL HLT 2011, Portland, June 2011, pp 107–115

Gao Yuqing, Bowen Zhou, Weizhong Zhu, Wei Zhang (2008) Handheld speech to speech translation system. Automatic speech recognition on mobile devices and over communication networks, Springer, London

Gildea D (2003) Loosely tree-based alignment for machine translation ACL-2003.In: 41st annual meeting of the association for computational linguistics, Sapporo, 7–12 July 2003

Gough N, Way A (2003) Controlled generation in example-based machine translation MT Summit IX, New Orleans, 23–27 Sept 2003, pp 133–140

Green T (1979) The necessity of syntax markers: two experiments with artificial languages. J Verbal Learn Behav 18:481–496

Habash N, Dorr B, Monz C (2009) Symbolic-to-statistical hybridization: extending generation-heavy machine translation. Mach Transl 23:23–63

Hanneman G, Lavie A (2009) Decoding with syntactic and non-syntactic phrases in a syntax-based machine translation system, 2009. In: Proceedings of SSST-3, third workshop on syntax and structure in statistical translation, ACL, pp 1–9

Haque R, Naskar SK, Ma Y, Way A (2009) Using supertags as source language context in SMT. In: Lluís Màrquez, Harold Somers (eds) EAMT-2009: proceedings of the 13th annual conference of the European association for machine translation, Universitat Politècnica de Catalunya, Barcelona, 14–15 May 2009, pp 234–241

He Xiaodong, Mei Yang, Jianfeng Gao, Patrick Nguyen, Robert Moore (2008) Indirect HMM based hypothesis alignment for combining outputs from machine translation systems. In:Proceedings of the 2008 conference on empirical methods in natural language processing, pp 98–107

Hermjakob Ulf (2009) Improved Word alignment with statistics and linguistic heuristics. In: Proceedings of EMNLP 2009, Singapore, pp 229–237

Heyn M (1996) Integrating machine translation into translation memory systems. In: Proceedings of the EAMT machine translation workshop, TKE '96, Vienna, pp 113–126

Hoang H, Koehn P (2010) Improved translation with source syntax labels. In: Proceedings of the joint 5th workshop on statistical machine translation and metrics MATR, Uppsala, 11–16 July 2010, pp 409–417

Hoang H, Koehn P, Lopez A (2009) A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In: IWSLT 2009: proceedings of the international workshop on spoken language translation, National Museum of Emerging Science and Innovation, Tokyo, 1–2 Dec 2009, pp 152–159

Hovy E (1998) Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In: Proceedings of the first international conference on language resources and evaluation, Granada

Hovy E, Nirenburg S (1992) Approximating and interlingua in a principled way. In: Proceedings of the DARPA speech and natural language workshop, Arden House

Hovy E, Marcus M, Palmer M, Pradhan S, Ramshaw I, Weischedel R (2006) OntoNotes: the 90 % solution. In: Proceedings of the human language technology conference at the annual meeting of NAACL, New York

Huang Liang, David Chiang (2007) Forest rescoring: faster decoding with integrated language models. In: ACL 2007: proceedings of the 45th annual meeting of the association for computational linguistics, Prague, June 2007, pp 144–151

Huang Liang, Knight K, Joshi A (2006) Statistical syntax-directed translation with extended domain of locality. In: AMTA 2006: proceedings of the 7th conference of the association for machine translation in the americas, visions for the future of machine translation, Cambridge, MA, 8–12 Aug 2006, pp 66–73

Huang Liang, Hao Zhang, Daniel Gildea, Kevin Knight (2009) Binarization of synchronous context-free grammars. Comput Linguist 35(4):559–595

Huang Chu-Ren, Ru-Yng Chang, Hsiang-bin Lee (2010) Sinica BOW (Bilingual Ontological WordNet): integration of bilingual WordNet and SUMO. In: Huang et al (eds) Ontology and the lexicon, CUP, Cambridge, pp 201–211

Huang Zhongqiang, Martin Cmejrek, Bowen Zhou (2010) Soft syntactic constraint for hierarchical phrase-based translation using latent syntactic distributions. In: Proceedings of EMNLP10, Cambridge, MA

Hutchins J (2005a) State of the art reports natural language translation computer-based translation systems and tools. www.hutchingsweb.me.uk/BCS-NLT-2005.pdf

Hutchins J (2005b) Current commercial machine translation systems and computer-based translation tools: system types and their uses. Int J Transl 17(1–2):5–38

Hutchins J (2010) Outline of machine translation developments in Europe and America. JAPIO, Tokyo, pp 1–8

Hutchins WJ, Somers HL (1992) An introduction to machine translation. Academic, London, Xxi, 362pp

Hwidong Na, Jong-Hyeok Lee (2011) Multi-word unit dependency forest-based translation rule extraction. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, ACL HLT 2011, pp 41–51

Hyoung-Gyu Lee, Min-Jeong Kim, Gumwon Hong, Sang-Bum Kim, Young-Sook Hwang, Hae-Chang Rim (2010) Identifying idiomatic expressions using phrase alignments in bilingual parallel corpus. In: Proceedings of PRICAI 2010, Daegu, Korea

Jelinek F (2004) Stochastic analysis of structured language modeling. In: Johnson M, Khudanpur S, Ostendorf M, Rosenfeld R (eds) Mathematical foundations of speech and language processing. Springer, Berlin, pp 37–72

Jianjun Ma, Degen Huang, Haixia Liu, Wenfeng Sheng (2011) POS tagging of English particles for machine translation. In: Proceedings of the 13th machine translation summit, Asia-Pacific Association for Machine Translation, Xiamen, pp 57–63

Jijkoun V, de Rijke M (2004) Enriching the output of a parser using memory-based learning. In: Proceedings of the 42nd annual meeting of the Association for Computational Linguistics

Johannessen JB, Nordgård T, Nygaard L (2008) Evaluation of linguistics-based translation. In: LREC 2008: 6th language resources and evaluation conference, Marrakech, pp 26–30

Johansson R, Nugues P (2007) Extended constituent-to-dependency conversion for english. In: Proceedings of NODALIDA 2007, Tartu

Josef OF, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Josef OF, Ney H (2004) The alignment template approach to statistical machine translation. Comput Linguist 30(4):417–449

Karakos D, Khudanpur S (2008) Sequential system combination for machine translation of speech. In: Proceedings of the 2008 IEEE workshop on spoken language technology (SLT-08), Goa

Kay M (2011) Zipf's Law and *L'Arbitraire du Signe*. Linguist Issues Lang Technol (LiLT) 6(8):1–25, CLSI Publications

Khadivi S, Zens R, Ney H (2006) Integration of speech to computer-assisted translation using finite-state automata. In: Proceedings of COLING/ACL 2006, Sydney

Khalilov M, Pretkalniņa L, Kuvaldina N, Pereseina V (2010) SMT of Latvian, Lithuanian and Estonian languages: a comparative study. In: Human language technologies – the Baltic perpective, international conference, Riga, 8 Oct 2010, 8pp

Kirchhoff K, Rambow O, Habash N, Diab M (2007) Semi-automatic error analysis for large-scale statistical machine translation systems. In: Proceedings of the machine translation summit (MT- Summit), Copenhagen

Klein D, Manning CD (2003) Accurate unlexicalized parsing. In: ACL 41, pp 423–430

Koehn P (2004) Statistical significance tests for machine translation evaluation. In: Proceedings of EMNLP, Barcelona, pp 388–395

Koehn P (2010) Statistical machine translation. Cambridge University Press, Cambridge, XII, 433p

Koehn P, Knight K (2003) Feature-rich statistical translation of noun phrases. In: Proceedings of ACL 2003, Hongkong

Koehn P, Monz C (2006) Manual and automatic evaluation of machine translation between European languages. In: NAACL 2006 workshop on statistical machine translation, ACL, New York, pp 102–121

Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of NAACL, Morristown, pp 48–54

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (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, pp 177–180

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Labaka G (2010) EUSMT: incorporating linguistic information into SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation. Ph.D. thesis, University of the Basque Country

Labaka G (2010) EUSMT: incorporating linguistic information into SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation. Ph.D. thesis, University of the Basque Country

Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML, pp 282–289

Lagarda A-L, Alabau V, Casacuberta F, Silva R, Díaz-de-Liaño E (2009) Statistical post-editing of a rule-based machine translation system. In: Proceedings of NAACL HLT 2009. Human language technologies: the 2009 annual conference of the North American chapter of the ACL, short papers, Boulder, 31 May–5 June 2009, pp 217–220

Lambert P, Banchs R (2005) Data inferred multi-word expressions for statistical machine translation. In: Proceedings of MT summit X, Phuket

Lamel L et al (2011) Speech recognition for machine translation in quaero. In: Proceedings of IWSLT, San Francisco

Lavie A (2008) Stat-XFER: a general search-based syntax-driven framework for machine translation. In: Computational linguistics and intelligent text processing, Springer, LNCS, New York, pp 362–375

Lavie A, Agarwal A (2007) METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the second workshop on statistical machine translation, Prague, pp 228–231

Lavie A, Waibel A et al (1996) Translation of conversational speech with JANUS-II. In: Proceedings of ICSLP 96, Philadelphia

Lavie A et al (2002) A multi-perspective evaluation of the NESPOLE! speech-to-speech translation system. In: Proceedings of ACL 2002 workshop on speech-tospeech translation: algorithms and systems, Philadelphia

Lavie A, Yarowsky D, Knight K, Callison-Burch C, Habash N, Mitamura T (2006) MINDS workshops machine translation working group final report. http://www-nlpir.nist.gov/MINDS/FINAL/MT.web.pdf

Lavie A, Parlikar A, Ambati V (2008) Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In: Proceedings of the second ACL workshop on syntax and structure in statistical translation, Columbus, pp 87–95

Lei Cui, Zhang D, Li M, Zhou M, Zhao T (2010) A joint rule selection model for hierarchical phrase-based translation. In: Proceedings of ACL10, Uppsala, pp 6–11

Lei Cui, Dongdong Zhang, Mu Li, Ming Zhou (2011) Function word generation in Statistical Machine Translation Systems, MTS 2011, pp 139–146

Lemao Liu, Tiejun Zhao, Chao Wang, Hailong Cao (2011) A unified and discriminative soft syntactic constraint model for hierarchical phrase-based translation

Levin L, Lavie A, Woszczyna M, Gates D, Galvadá M, Koll D, Waibel A (2000) The janus-III translation system: speech-to-speech translation in multiple domains. Mach Trans 15:3–25

Levy R, Manning C (2004) Deep dependencies from context–free statistical parsers: correcting the surface dependency approximation. In: Proceedings of the ACL

Li Zhifei, Callison-Burch C, Dyer C, Ganitkevitch J, Khudanpur S, Schwartz L, Thornton WNG, Weese J, Zaidan OF (2009) Demonstration of Joshua: an open source toolkit for parsing-based machine translation. In: Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP, Software demonstrations, Suntec, 3 Aug 2009, pp 25–28

Liu Ding, Gildea D (2008) Improved tree-to-string transducer for machine translation. In: Proceedings of ACL-08: HLT. Third workshop on statistical machine translation (ACL WMT-08), The Ohio State University, Columbus, 19 June 2008, pp 62–69

Liu Ding, Gildea D (2010) Semantic role features for machine translation. In: Proccedings of the coling 2010: 23rd international conference on computational linguistics, Beijing International Convention Center, Beijing, 23–27 Aug 2010, pp 716–724

Liu Yang, Qun Liu, Shouxun Lin (2006) Tree-to-string alignment template for statistical machine translation. In: Coling-ACL 2006: proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, Sydney, 17–21 July 2006, pp 609–616

Liu Yang, Yajuan Lü, Qun Liu (2009) Improving tree-to-tree translation with packed forests. In: Proceedings of the 47th annual meeting of the ACL and the 4th IJCNLP, Suntec, 2–7 Aug 2009, pp 558–566

Liu Zhanyi, Haifeng Wang, Hua Wu, Sheng Li (2010) Improving statistical machine translation with monolingual collocation. In: Proceedings of ACL 2010, Uppsala

Liu Lemao, Tiejun Zhao, Chao Wang, Hailong Cao (2011a) A unified and discriminative soft syntactic constraint model for hierarchical phrase-based translation

Liu Shujie, Chi-Ho Li, Ming Zhou (2011b) A unified SMT framework combining MIRA and MERT in MTS, pp 181–188

Liu Yang, Qun Liu, Yajuan Lü (2011) Adjoining tree-to-string translation. In: ACL-HLT 2011: proceedings of the 49th annual meeting of the association for computational linguistics, Portland, 19–24 June 2011, pp 1278–1287

Llitjós AF, Vogel S (2007) A walk on the other side. Adding statistical components to a transfer-based translation system. In: Proceedings of the HLT-NAACL workshop on syntax and structure in statistical translation, Rochester, pp 72–79

Lopez A, Resnik P (2006) Word-based alignment, phrase-based translation: what's the link. In: Proceedings of AMTA, pp 90–99

Lopez A, Resnik P (2006) Word-based alignment, phrase-based translation: what's the link. In: Proceedings of the AMTA, Cambridge MA, pp 90–99

Maletti A (2010) Why synchronous tree substitution grammars? In: Proceedings of the 2010 meeting of the North American chapter of the association for computational linguistics (NAACL-10), pp 876–884

Maletti A, Graehl J, Hopkins M, Knight K (2009) The power of extended top-down tree transducers. SIAM J Comput 39:410–430

Marcu D, Wei Wang, Echihabi A, Knight K (2006) SPMT: statistical machine translation with syntactified target language phrases. In: Proceedings of EMNLP 2006, pp 44–52

Mareček D (2009a) Improving word alignment using alignment of deep structures. In: Proceedings of the 12th international conference on text, speech and dialogue, pp 56–63

Mareček D (2009b) Using tectogrammatical alignment in phrase-based machine translation. In: Proceedings of WDS 2009 contributed papers, pp 22–27

Marton Y, Resnik P (2008) Soft syntactic constraints for hierarchical phrased-based translation. Proceedings of the ACL-08: HLT 46th annual meeting of the association for computational linguistics: human language technologies, The Ohio State University, Columbus, 15–20 June 2008, pp 1003–1011

Maruyama H, Watanabe H (1992) Tree cover search algorithm for example-based translation fourth international conference on theoretical and methodological issues in machine translation (TMI-92), empiricist vs. rationalist methods in MT, Montreal, CCRIT-CWARC, 25–27 June 1992, pp 173–184

Matthew GS, Madnani N, Dorr B, Schwartz R (2009) TER-Plus: paraphrase, semantic, and alignment enhancements to translation error rate. Mach Trans 23(2/3):117–127

McCrae J, Campana J, Cimiano P (2010) CLOVA: an architecture for cross-language semantic data querying. In: Proceedings of the first mutlilingual semantic web workshop

McCrae et al (2011) Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In: Proceedings of SSST, pp 116–125

McCrae J, Espinoza M, Montiel-Ponsoda E, Aguado-de-Cea G, Cimiano P (2011) Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, ACL HLT 2011, Portland, pp 116–125

Menezes A, Quirk C (2008) Syntactic models for structural word insertion and deletion. In: Proceedings of EMNLP

Menezes A, Richardson SD (2001) A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In: Proceedings of MT summit VIII workshop on example-based machine translation, Santiago de Compostela, 18–22 Sept 2001

Menezes A, Toutanova K, Quirk C (2006) Microsoft research treelet translation system: NAACL 2006 Europarl evaluation. In: HLT-NAACL 2006: proceedings of the workshop on statistical machine translation, New York, June 2006, pp 158–161

Mi Haitao, Liang Huang, Qun Liu (2008) Forest-based translation. In: Proceedings of the ACL-08: HLT: 46th annual meeting of the association for computational linguistics: human language technologies, 15–20 June 2008, The Ohio State University, Columbus, pp 192–199

Michael C, Way A (eds) (2003) Recent advances in example-based machine translation. Kluwer Academic, Dordrecht. Introduction to the workshop on EBMT, xxxi

Ming Tan, Wenli Zhou, Lei Zheng, Shaojun Wang (2012) A scalable distributed syntactic, semantic and lexical language model, to appear in computational linguistics just accepted MS, pp 1–66

Mischo W (1982) Library of congress subject headings. Catalog Classif Quart 1(2):105–124

Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (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, pp 177–180

Morimoto T et al ATR's speech translation system: ASURA. In: Proceedings of EuroSpeech 93, pp 1291–1294

Na Hwidong, Lee Jong-Hyeok (2011) Multi-word unit dependency forest-based translation rule extraction. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, ACL HLT 2011, pp 41–51

Nakazawa T, Kurohashi S (2008) Linguistically-motivated tree-based probabilistic phrase alignment. In: AMTA-2008 MT at work: proceedings of the eighth conference of the association for machine translation in the Americas, Waikiki, 21–25 Oct 2008, pp 163–171

Nakazawa T, Kurohashi S (2011) Statistical phrase alignment model using dependency relation probability. In: Proceedings of SSST-3, third workshop on syntax and structure in statistical translation, pp 10–18

Navigli R, Ponzetto SP (2010) Babelnet: building a very large multilingual semantic network. In: Proceedings of the 48th annual meeting of the association for computational linguistics, pp 216–225

Nießen S, Och FJ, Leusch G, Ney H (2000) An evaluation tool for machine translation: fast evaluation for MT research. In: Proceedings of LREC-2000: second international conference on language resources and evaluation, Athens, 31 May–2 June 2000, pp 39–45

Och FJ (2003) Minimum error rate training in statistical machine translation. In: Proceedings of ACL, pp 160–167

Och FJ, Ney H (2002) Discriminative training and maximum entropy models for statistical machine translation. In: ACL 2002: proceedings of the 40th annual meeting of the association for computational linguistics (best paper award), Philadelphia, July 2002, pp 295–302

Och FJ, Ney H (2003a) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Och FJ, Ney H (2003b) The alignment template approach to statistical machine translation. Comput Linguist 30(4):417–449

Oepen S, Lønning JT (2006) Discriminant-based MRS banking. In: Proceedings of the 4th international conference on language resources and evaluation

Oepen S, Dyvik H, Lønning JT, Velldal E, Beermann D, Carroll J, Flickinger D, Hellan L, Johannessen JB, Meurer P, Nordgård T, Rosén V (2004) Som a kapp-ete med trollet? Towards MRS-based

Norwegian–english machine translation. In: Proceedings of the 10th international conference on theoretical and methodological issues in machine translation, Baltimore, pp 11–20
Oepen S, Dyvik H, Flickinger D, Lønning JT, Meurer P, Rosén V (2005) Holistic regression testing for high-quality MT. Some methodological and technological reflections. In: Proceedings of the 10th annual conference of the European association for machine translation, Budapest
Oepen S, Velldal E, Lønning JT, Meurer P, Rosén V, Flickinger D (2007) Towards hybrid quality-oriented machine translation—on linguistics and probabilities in MT. In: TMI-2007: proceedings of the 11th international conference on theoretical and methodological issues in machine translation, Skövde, 7–9 Sept 2007, pp 144–153
Oflazer K, El-Kahlout ID (2007) Exploring different representational units in English-to-Turkish statistical machine translation. In: Proceedings of the second workshop on statistical machine translation, ACL, pp 25–32
Owczarzak K, van Genabith J (2007) Evaluating machine translation with LFG dependencies [abstract]. Mach Trans 21(2):95–119
Papineni K, Roukos S, Ward T, Wei-jing Zhu (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL
Pekar V, Mitkov R, Blagoev D, Mulloni A (2006) Finding translations for low-frequency words in comparable corpora. Mach Transl 20(4):247–266
Petrov S, Barrett L, Thibaux R, Klein D (2006) Learning accurate, compact, and interpretable tree annotation. In: Proceedings of COLING-ACL
Phillips AB (2011) Cunei: open-source machine translation with relevance-based models of each translation instance, in special issue: free/open-source machine translation machine translation. Mach Trans 25(2):161–177
Philpot A, Hovy E, Pantel P (2010) The OMEGA ontology. In: Huang CR et al (eds) Ontology and the lexicon. Cambridge University Press, Cambridge, UK, pp 258–270
Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, Manning CD (2011) Discriminative reordering with Chinese grammatical relations features. In: Proceedings of SSST-3, third workshop on syntax and structure in statistical translation, Boulder, pp 51–59
Pighin Daniele, Lluìs Màrquez (2011) Automatic projection of semantic structures: an application to pairwise translation Popović, Maja & Hermann Ney: 2006. POS-based reorderings for statistical machine translation. In: LREC-2006: fifth international conference on language resources and evaluation, Genoa, 22–28 May 2006, pp1278–1283
Pighin D, Màrquez L (2011) Automatic projection of semantic structures: an application to pairwise translation ranking. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical rranslation, Portland, pp 1–9
Przybocki M, Peterson K, Bronsart S (2008) Translation adequacy and preference evaluation tool (TAP-ET). In: LREC 2008: 6th language resources and evaluation, Marrakech, 26–30 May 2008, 8pp
Qin Gao, Vogel S (2011) Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, ACL HLT 2011, Portland, pp 107–115
Quirk C, Menezes A, Cherry C (2005) Dependency treelet translation: syntactically informed phrasal SMT. In: ACL-2005: 43rd annual meeting of the association for computational linguistics, University of Michigan, Ann Arbor, 25–30 June 2005, pp 271–279
Ranking, in proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, pp 1–9
Ravi S, Knight K (2010) Does GIZA++ make search errors? Comput Linguist Squibs Discuss 36(3):295–302
Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang (2009) Improving statistical machine translation using domain bilingual multiword expressions. In: Proceedings of MWE 2009 (ACL-IJCNLP)
Riezler S, Maxwell JT III (2006) Grammatical machine translation. In: Proceedings of the human language technology conference and annual meeting of the North American association for computational linguistics, pp 248–255

Rosti AVI, Bing Xiang, Matsoukas S, Schwartz R, Ayan NF, Dorr BJ (2007a) Combining outputs from multiple machine translation systems. In: Proceedings of HLT-NAACL, Rochester, pp 228–235

Rosti AVI, Matsoukas S, Schwartz R (2007b) Improved word-level system combination for machine translation. In: Proceedings of ACL-07, pp 312–319

Rosti AVI, Bing Zhang, Matsoukas S, Schwartz R (2008) Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In: Proceedings of ACL/WMT 2008, pp 183–186

Roturier J (2009) Deploying novel MT technology to raise the bar for quality: a review of key advantages and challenges. Keynote slides, the twelfth machine translation summit, International association for machine translation, Ottawa

Saers M (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, Uppsala

Saers M, Nivre J, Wu D (2010) Word alignment with stochastic bracketing linear inversion transduction grammar. In: HLT/NAACL2010, ACL, pp 341–344

Sag IA, Baldwin T, Bond F, Copestake A, Flickinger D (2001) Multiword expressions: a pain in the neck for NLP. In: Proceedings of WP

Sanchez-Cartagena VM, Pérez-Ortiz JA (2010) Tradubi: open-source social translation for the Apertium machine translation platform. In: Open source tools for machine translation, MT Marathon 2010, pp 47–56

Sanchez-Cartagena VM, Sanchez-Martìnez F, Pérez-Ortiz JA (2011) Integrating shallow-transfer rules into phrase-based statistical machine translation, MT Summit XIII: the thirteenth machine translation summit [organized by the] Asia-Pacific association for machine translation (AAMT), Xiamen, 19–23 Sept 2011, pp 562–569

Sanchez-Martinez F, Forcada ML (2009) Inferring shallow-transfer machine translation rules from small parallel corpora. J Artif Intell Res 34:605–635

Schafer C, David Y (2003) Statistical machine translation using coercive two-level syntactic transduction EMNLP-2003. In: proceedings of the 2003 conference on empirical methods in natural language processing, a meeting of SIGDAT, a special interest group of the ACL, held in conjunction with ACL-03, Sapporo, 11–12 July 2003, 8pp

Schafer C, Yarowsky D (2002) Inducing translation lexicons via diverse similarity measures and bridge languages. In: CoNLL, Taipei

Schmid H (2006) Trace prediction and recovery with unlexicalized PCFGs and slash features. In Proceedings of the COLING-ACL, Sydney

Schwenk H, Abdul-Rauf S, Barrault L, Senellart J (2009) SMT and SPE machine translation system for WMT'09. In: Proceedings of the fourth workshop on statistical machine translation, Athens, 30–31 March 2009, pp 130–134

Shen Libin, Jinxi Xu, Weischedel R (2008) A new string-to-dependency machine translation algorithm with a target dependency language model. ACL-08: HLT. In: 46th annual meeting of the association for computational linguistics: human language technologies. Proceedings of the conference, The Ohio State University, Columbus, 15–20 June 2008, pp 577–585

Shen Libin, Jinxi Xu, Bing Zhang, Matsoukas S, Weischedel R (2009) Effective use of linguistic and contextual information for statistical machine translation. EMNLP-2009. In: Proceedings of the 2009 conference on empirical methods in natural language processing, Singapore, 6–7 Aug 2009, pp 72–80

Shu Cai, Chiang D, Goldberg Y (2011) Language-independent parsing with empty elements. In: Proceedings of the 49th annual meeting of the ACL, Portland, pp 212–216

Shujie Liu, Chi-Ho Li, Ming Zhou (2011) A unified SMT framework combining MIRA and MERT in MTS, pp 181–188

Shumin Wu, Palmer M (2011) Semantic mapping using automatic word alignment and semantic role labeling. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, Portland, pp 21–30

Shumin Wu, Choi JD, Palmer M (2010) Detecting cross-lingual semantic similarity using parallel propbanks. In: Proceedings of the 9th conference of the association for machine translation in the Americas

Sim, Khe Chai, William JB, Mark JFG, Hichem S, Phil CW (2007) Consensus network decoding for statistical machine translation system combination. In: Proceedings of the 32nd IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 105–108

Simard M, Ueffing N, Isabelle P, Kuhn R (2007) Rule-based translation with statistical phrase-based post-editing. In: Proceedings of the second workshop on statistical machine translation, ACL, pp 203–206

Simard M, Cyril G, Pierre I (2007) Statistical phrase-based post-editing. NAACL-HLT-2007 Human language technology: the conference of the North American chapter of the association for computational linguistics, Rochester, 22–27 April 2007, pp 508–515

Smith DA, Eisner J (2006) Minimum risk annealing for training log-linear models. In: Proceedings of the COLING/ACL on main conference poster sessions, ACL, pp 787–794

Snover M, Bonnie D, Richard S, Linnea M, John M (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the association for machine translation in the Americas (AMTA-2006), Cambridge, MA, Aug, pp 223–231

Soltau H, Metze F, Fügen C, Waibel A (2001) A one pass-decoder based on polymorphic linguistic context assignment. In: IEEE ASRU, Madonna di Campiglio

Somers H (2003) An overview of EBMT. In: Michael C, Andy W (eds) Recent advances in example-based machine translation. Kluwer Academic, Dordrecht, pp 3–57

Specia L, Cancedda N, Turchi M, Cristianini N (2009) Estimating the sentence-level quality of machine translation systems. In: Proceedings of the 13th annual conference of the EAMT, pp 28–35

Specia L, Saunders C, Turchi M, Wang Z, Shawe-Taylor J (2009) Improving the confidence of machine translation quality estimates. MT summit XII

Stein D, Stephan P, David V, Hermann N (2010) A cocktail of deep syntactic features for hierarchical machine translation. AMTA 2010: the ninth conference of the association for machine translation in the Americas, Denver 31 Oct–4 Nov 2010, 9pp

Stolcke A (2002) SRILM – an extensible language modeling toolkit. In: Proceedings of the international conference of spoken language processing, vol 2, Denver, pp 901–904

Suzuki H (2011) Automatic post-editing based on SMT and its selective application by sentence-level automatic quality evaluation, MTS, pp 156–163

Tan Ming, Wenli Zhou, Lei Zheng, Shaojun Wang (2012) A scalable distributed syntactic, semantic and lexical language model, to appear in computational linguistics Just accepted MS, pp 1–66

Teramusa E (2007) Rule based machine translation combined with statistical post-editor for Japanese to English patent translation. Tokyo University of Science, Suwas

Thurmair G (2009) Comparing different architectures of hybrid machine translation systems. In: Proceedings of MT summit XII

Tiedemann J (2011) Bitext alignment. Morgan & Claypool Publishers, viii, 153 p

Tiedemann J, Kotzé G (2009) Building a large machine-aligned parallel treebank. In: Proceedings of the 8th international workshop on treebanks and linguistic theories (TLT), Milan, pp 197–208

Tillmann C (2004) A unigram orientation model for statistical machine translation. HLT-NAACL 2004: Human language technology conference and North American chapter of the association for computational linguistics annual meeting, The Park Plaza Hotel, Boston, – Short Papers, 2–7 May 2004, pp 101–104

Tinsley J, Hearne M, Way A (2007) Exploiting parallel treebanks to improve phrase-based statistical machine translation. In: Proceedings of the sixth international workshop on treebanks and linguistic theories, pp 175–187

Tonelli S, Delmonte R, Bristot A (2008) Enriching the Venice Italian Treebank with dependency and grammatical relations. In: Proceedings of LREC 2008, Marrakech

Tonelli S, Rodolfo D, Antonella B (2008) Enriching the Venice Italian treebank with dependency and grammatical relations, LREC 2008

Tong Xiao, Jingbo Zhu, Shujie Yao, Hao Zhang (2011) Document-level consistency verification in machine translation, MST 2011, pp 131–138

Turian JP, Luke S, Melamed ID (2003) Evaluation of machine translation and its evaluation MT Summit IX, New Orleans, 23–27 Sept 2003, pp 386–393

Tyers FM (2009) Rule-based augmentation of training data in Breton-French statistical machine translation. In: Proceedings of the 13th annual conference of the European association for machine translation, pp 213–217

Ueffing N, Haffari G, Sarker A (2007) Semi-supervised model adaptation for statistical machine translation. Mach Trans 21(2):77–94

Vandeghinste V, Scott M (2010) Bottom-up transfer in example-based machine translation. In: François I, Veale, T, Andy W (eds) 1997 Gaijin: a bootstrapping, template-driven approach to example-based MT. Proceedings of the 2nd international conference on recent advances in natural language processing, Tzigov Chark1997, pp 239–244

Vandeghinste V, Van den Bogaert J, Martens S, Kotzé G (2011) PaCo-MT: parse and corpus-based machine translation. In: Forcada ML, Depraetere H, Vandeghinste V (eds) Proceedings of the 15th annual conference of the European association for machine translation, p 347

Velldal E, Oepen S (2006) Statistical ranking in tactical generation. In: Proceedings of the conference on empirical methods in natural language processing. Sydney

Velldal E, Oepen S, Flickinger D (2004) Paraphrasing treebanks for stochastic realization ranking. In: Proceedings of the 3rd workshop on treebanks and linguistic theories, pp 149–160

Venkatapathy S, Sangal R, Joshi A, Gali K (2010) A discriminative approach for dependency based statistical machine translation. In: Proceedings of SSST, pp 66–74

Venugopal A, Andreas Z, Noah AS, Stephan V (2009) Preference grammars: softening syntactic constraints to improve statistical machine translation. NAACL HLT 2009. Human language technologies: the 2009 annual conference of the North American chapter of the ACL, Boulder, 31 May–5 June 2009, pp37–45

Viggo H (eds) Proceedings of the 14th international conference of the European association for machine translation (EAMT-2010), 8pp

Vossen P (1998) EuroWordNet: a multilingual database with lexical semantic networks. Kluwer, Dordrecht

Waibel A, Ahmed B, Alan WB, Robert F, Donna Gates AL, Lori L, Kevin L, Laura MT, Juergen R, Tanja S, Dorcas W, Monika W, Jing Z (2003) Speechalator: two-way speech-to-speech translation in your hand. In: Proceedings of HLT-NAACL 2003, Demonstrations, May–June 2003, pp 29–30

Wallach H (2006) Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on machine learning (ICML), New York, pp 977–984

Wang S, Wang S, Greiner R, Schuurmans D, Cheng L (2005) Exploiting syntactic, semantic and lexical regularities in language modeling via directed Markov random fields. In: Proceedings of the 22nd international conference on machine learning (ICML), Bonn, pp 953–960

Wang S, Wang S, Cheng L, Greiner R, Schuurmans D (2006) Stochastic analysis of lexical and semantic enhanced structural language model. In: Proceedings of the 8th international colloquium on grammatical inference (ICGI), Tokyo, pp 97–111

Wang Chao, Michael Collins, Philipp K (2007) Chinese syntactic reordering for statistical machine translation. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), Prague, pp 737–745

Wang Wei, Kevin Knight, Daniel Marcu (2007) Binarizing syntax trees to improve syntax-based machine translation accuracy. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning, Prague, pp 746–754

Wei Wang, May J, Knight K, Marcu D (2010) Re-structuring, re-labeling, and re-aligning for syntax-based machine translation. Comput Linguist 36(2):247–277

Watanabe Hideo, Sadao Kurohashi, Eiji Aramaki (2000) Finding structural correspondences from bilingual parsed corpus for corpus-based translation Coling 2000 in Europe: the 18th international conference on computational linguistics. In: Proceedings of the conference, Universität des Saarlandes, Saarbrücken, 31 July–4 Aug 2000, pp 906–912

Wehrli E (2007) Fips, a "deep" linguistic multilingual parser. In: Proceedings of the ACL 2007 workshop on deep linguistic processing, Prague, pp 120–127

Wehrli E, Nerima L, Seretan V, Scherrer Y (2009) On-line and off-line translation aids for non-native readers. In: Proceedings of the international multiconference on computer science and information technology, vol 4, pp 299–303

Wehrli E, Seretan V, Nerima L, Russo L (2009) Collocations in a rule-based MT system: a case study evaluation of their translation adequacy. In: EAMT-2009: proceedings of the 13th annual conference of the European association for machine translation, Lluís Màrquez, Harold Somers (eds), 14–15 May 2009, Universitat Politècnica de Catalunya, Barcelona, pp 128–135

Wei Wang, Knight K, Marcu D (2007) Binarizing syntax trees to improve syntax-based machine translation accuracy. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning, pp 746–754

Wilks Y (1994) Stone soup and the French room. In: Zampolli A, Calzolari N, Palmer M (eds) Current issues in computational linguistics: in honour of Don Walker, vol 9–10, Linguistica Computazionale. Giradini Editori/Kluwer Academic, Pisa/Dordrecht, pp 585–594

Wilks Y (2009) Machine translation: its scope and limits. Springer, New York, 252 p

Wong Fai, Dong-Cheng Hu, Yu-Hang Mao, Ming-Chui Dong, Yi-Ping Li (2005) Machine translation based on constraint-based synchronous grammar. In: Proceedings of the 2nd international joint conference on natural language, Jeju Island, pp 612–623

Wu D (1997) Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Comput Ling 23(3):378–403

Wu Dekai, Hongsing Wong (1998) Machine translation with a stochastic grammatical channel. In: Coling-ACL '98: 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics, Université de Montréal, Montreal, 10–14 Aug 1998, pp 1408–1415

Wu Shumin, Martha Palmer (2011) Semantic mapping using automatic word alignment and semantic role labeling. In: Proceedings of SSST-5, fifth workshop on syntax, semantics and structure in statistical translation, Portland, pp 21–30

Wu Dekai, Pascale Fung (2009) Semantic roles for SMT: a hybrid two-pass model. In: NAACL HLT 2009: human language technologies: the 2009 annual conference of the North American chapter of the ACL, Short papers, Boulder, 31 May–5 June 2009, pp 13–16

Wu Shumin, Jinho D Choi, Martha Palmer (2010) Detecting cross-lingual semantic similarity using parallel propbanks. In: Proceedings of the 9th conference of the association for machine translation in the Americas

Wuebker J, Mauser A, Ney H (2010) Training phrase translation models with leaving-one-out. In: Proceeding of ACL, pp 475–484

Tong Xiao, Jingbo Zhu, Shujie Yao, Hao Zhang (2011) Document-level consistency verification in machine translation. MST 2011:131–138

Xiong Deyi, Qun Liu, Shouxun Lin (2006) Maximum entropy based phrase reordering model for statistical machine translation. In: Coling-ACL 2006: proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics, Sydney, 17–21 July 2006, pp 521–528

Xiong D, Zhang M, Li H (2010) Error detection for statistical machine translation using linguistic features. In: ACL 2010: the 48th annual meeting of the association for computational linguistics, Uppsala, pp 604–611

Xiong Deyi, Min Zhang, Haizhou Li (2010a) Learning translation boundaries for phrase-based decoding. In: Proceedings of HLT-NAACL 2010

Yamabana Kiyoshi et al (2003) A speech translation system with mobile wireless clients. In: Proceedings of ACL 2003

Yamada Kenji, Kevin Knight (2001) A syntax-based statistical translation model ACL-EACL-2001: 39th annual meeting [of the association for computational linguistics] and 10th conference of the European chapter (of ACL), Toulouse, 9–11 July 2001, pp 523–530

Yamada Kenji, Kevin Knight (2002) A decoder for syntax-based statistical MT. In: ACL-2002: 40th annual meeting of the association for computational linguistics, Philadelphia, July 2002, pp 303–310. (PDF, 788 KB)

Yaqin Yang, Nianwen Xue (2010) Chasing the ghost: recovering empty categories in the Chinese Treebank. In: Proceedings of COLING, Beijing

Yonggang Deng, Jia Xu, Yuqing Gao (2008) Phrase table training for precision and recall: what makes a good phrase and a good phrase pair? In: Proceedings of ACL, Columbus, pp 81–88

Zens Richard, Hermann Ney (2006) Discriminative reordering models for statistical machine translation. In: HLT-NAACL 2006: proceedings of the workshop on statistical machine translation, New York, June 2006, pp 55–63

Zhai Feifei, Jiajun Zhang Yu Zhou, Chengqing Zong (2011) Simple but effective approaches to improving tree-to-tree model, MTS

Zhang Y (2008) Structured language models for statistical machine translation. Ph.D. dissertation, Carnegie Mellon University

Zhang Licheng Fang, Peng Xu, Xiaoyun Wu (2011) Binarized forest to string translation. In: ACL-HLT 2011: proceedings of the 49th annual meeting of the association for computational linguistics, Portland, 19–24 June 2011, pp 835–845

Zhang Ying (Joy) http://projectile.sv.cmu.edu/research/public/talks/speechtranslation/sst-survey-joy.pdf

Zhang Ying, Stephan Vogel (2007) PanDoRA: a large- scale two-way statistical machine translation system for hand-held devices. In: Proceedings of MT Summit XI, Copenhagen, pp 10–14

Zhang Min, Hongfei Jiang, Ai Ti Aw, Jun Sun, Sheng Li, Chew Lim Tan (2007) A tree-to-tree alignment-based model for statistical machine translation. In: Proceedings of MT Summit XI, Copenhagen, 10–14 Sept 2007, pp 535–542

Zhang Min, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, Sheng Li (2008) A tree sequence alignment-based tree-to-tree translation model. In: Proceedings of the conference ACL-08: HLT. 46th annual meeting of the association for computational linguistics: human language technologies, The Ohio State University, Columbus, 15–20 June 2008, pp 559–567

Zhanyi Liu, Haifeng Wang, Hua Wu, Sheng Li (2010) Improving statistical machine translation with monolingual collocation. In: Proceedings of ACL 2010

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, Yun Huang (2009) Improving statistical machine translation using domain bilingual multiword expressions. In: Proceedings of MWE 2009 (ACL-IJCNLP)

Zhongqiang Huang, Cmejrek M, Zhou B (2010) Soft syntactic constraint for hierarchical phrase-based translation using latent syntactic distributions. In: Proceedings of EMNLP10, Stroudsburg

Zollmann A, Venugopal A (2006) Syntax augmented machine translation via chart parsing. In: Proceedings of the workshop on statistical machine translation, New York, pp 138–141

Zong Chenqing, Bo Xu, Taiyi Huang (2002) Interactive chinese-to-english speech translation based on dialogue management. In: Proceedings of the workshop on speech-to-speech translation: algorithms and systems, pp 61–68

Zong Chengqing, Heyan Huang, Shuming Shi (2008) Application of machine translation during Olympic Games 2008. In: AMTA-2008. MT at work: proceedings of the eighth conference of the association for machine translation in the Americas, Waikiki, 21–25 Oct 2008, pp 470–479

Online MT Systems and Tools

- http://www.languageweaver.com
- http://translate.google.com
- http://www.microsofttranslator.com
- http://www.systran.co.uk/
- http://www.freetranslation.com
- https://apertium.svn.sourceforge.net/svnroot/apertium/branches/apertium-dixtools-paradigmlearning
- http://www.opentrad.com
- http://www.eitb24.com/en
- http://www.cunei.org
- http://www.undl.org/
- http://www.unlweb.net/wiki/index.php/Introduction_to_UNL
- http://speechtrans.com/#
- http://researcher.ibm.com/researcher/view_page.php?id=2323
- http://www.youtube.com/watch?v=Ex-NBO_w0zQ
- http://www.babylon.com/mac.html
- http://itunes.apple.com/us/app/voicetra-speech-to-speech/id383542155?mt=8
- http://text-to-speech.imtranslator.net/
- http://www.speech.sri.com/projects/translation/
- http://www.bbn.com/technology/speech/speech_to_speech_translation
- http://www.ustar-consortium.com/
- http://www.research.att.com/projects/Speech_Translation/index.html?fbid=0GMC-dWS68d
- http://www.gizmag.com/go/2686/
- http://www.quaero.org
- http://www.speech.cs.cmu.edu/
- http://www.loquendo.com/it/
- http://www.is.cs.cmu.edu/mie/janus.html
- http://www-01.ibm.com/software/websphere/products/mobilespeech/
- http://research.microsoft.com/en-us/groups/srg/default.aspx
- http://tldp.org/HOWTO/Speech-Recognition-HOWTO/software.html
- http://cmusphinx.sourceforge.net/
- http://htk.eng.cam.ac.uk/
- http://julius.sourceforge.jp/en_index.php
- http://www.simon-listens.org/index.php?id=122&L=1
- http://www.sdl.com/en/language-technology/products/automated-translation/
- http://logos-os.dfki.de/
- http://www.openmatrex.org/
- http://tool.statmt.org/
- http://www.apertium.org/
- www.limsi.fr/tlp
- www.informatik.kit.edu/interact
- www-i6.informatik.rwth-aachen.de
- www.vocapia.com
- http://www.darpa.mil/ipto/Programs/gale/index.htm
- http://www.darpa.mil/
- http://www.ldc.upenn.edu/
- https://wit3.fbk.eu/
- http://iwslt2012.org/
- http://iwslt2012.org/index.php/evaluation-campaign/ted-task