Tree-based hybrid machine translation

Andreas Søeborg Kirkedal

Supervised by Professor Hans Uszkoreit and
Dr. Patrick Blackburn(Université Nancy 2)

UNIVERSITÄT DES SAARLANDES

August 4, 2011

Abstract

I present a post-editing approach that combines translation systems which produce syntactic trees as output. The nodes in the generation tree and target side hierachical tree are aligned and form the basis for computing structural similarity. Structural similarity computation aligns subtrees and based on this alignment, subtrees are substituted to create more accurate translations. Two different techniques have been implemented to compute structural similarity: Leaves and Tree Edit distance. I report on the translation quality of two hybrid MT systems where both techniques are implemented. The approach shows significant improvement over the baseline for MT systems trained with limited training data and structural improvement for MT systems trained on Europarl.

Declaration

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, 05 August 2011

Signature:

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many thousands of machine-readable documents are produced every day in hundreds of different languages. With the advent of the web, most of the documents are accessible. However, not everyone can understand the information made accessible on the web. A recent Eurobarometer survey carried out by Gallup in 27 European countries found that 90% of internet users prefer their own language when surfing the web and that 44% feel that they are missing interesting information because it is not in their own language [Eurobarometer, 2011]. The language barrier is one of the last great barriers to the free flow of information.

This is especially a problem for small languages. Languages that do not have a large number of speakers, such as Danish, will not have the same information available to Danes as is available to English speakers. The information may only be available in English or Chinese and with the sheer number of new documents that are produced and the lack of a financial incentive to translate documents on the web and subsequently make them freely available, machine translation remains the only alternative method to access this information.

Using machine translation, a user can access information in a different language on demand as long as the software is available for the language pair in question.

1.1 History of machine translation

The idea of using computers to translate from one language into another was proposed as early as 1949 [Weaver, 1955]. Weaver compared translation to cryptography. The comparison is not accurate as both languages are English and are explicitly linked to each other by a set of operations. This is not the case in machine translation, as the natural languages involved evolve independently of each other.

Instead of using a statistical approach as envisioned by Weaver, the first machine translation (MT) systems relied on a rule-based approach. In this approach, linguists write rules that analyse the source language and can create an intermediate representation of language. Another set of linguistic rules generate a translation from this interpretation. The intention is to create an abstract, language-independent representation of the meaning of language and this representation is called an *interlingua representation*. Intuitively, this approach is similar to how many people think translation is performed and if an interlingua representation was discovered, a translation system would only have to create this deep analysis of a language and could then generate any target language from it. So far, it has not been possible to create an interlingua representation that was genuinely language-independent [Hutchins and Somers, 1992] and other intermediate structures have been proposed. The rule-based machine translation (RBMT) systems that rely on a less abstract interpretation are called *transfer-based* because a set of rules is needed to transfer the information in the intermediate structures — which are language-dependent — into intermediate structures of the target language.

In the beginning, the primary sponsors of MT research were military and intelligence agencies in the US who were interested in translating Russian. They ordered the ALPAC report which virtually ended MT research in the US. It concluded that MT was slower, less accurate and more expensive than using human translators [Hutchins, 2007]. The interlingua model was criticised as being too rigid and the accuracy of the analyses at lower linguistic levels was not high enough to make the interlingua approach robust. Research in MT did not gain momentum again until the mid 1970s and started focusing on transfer-based approaches.

Statistical machine translation (SMT) was developed in the early 1990s. The availability of corpora and increased computing power made it feasible to use statistics to compute translation probabilities and create probabilistic translation models. The research was initiated by scientists at IBM [Brown et al., 1990]. The development of SMT has revived MT research, gained a large momentum and reached the same level of quality as RBMT and in many cases outperforms rule-based systems.

1.2 State-of-the-art

Currently, state-of-the-art systems use a statistical approach to translation. Machine learning algorithms are applied to parallel corpora of previously translated text and the system will then be able to translate unseen text. Using a SMT toolkit and a parallel corpus, a statistical system can be built in a few weeks for a language pair.

The quality of the output of such a system is sensitive to the amount of available data. Parallel data is not as common as monolingual data and therefore focus has been on increasing the training data of the component that needs monolingual data - the language model. Google created the giga n-gram language model dataset [Brants and Franz, 2006] by crawling the web. A data set extracted from books in Google Books to increase the amount of parallel data to train translation models was also compiled, but while translations did improve, the improvement ratio relative to the extra amount of data has started to increase, indicating that a different approach might be needed instead of just adding more data.

Adding linguistic information to the statistical approach is a research area that also receives a lot of attention. This approach is also the only recourse for small or low-resource languages that do not have huge amounts of data available. A system that uses linguistic information will be better at generalising over phrases and can compensate for the lack of training data.

Several methods for introducing linguistic information have been proposed, e.g. factored models, hierarchical phrases and syntax-based models. Factored models create a vector of tags for each word, while hierarchical phrases take into account the recursive nature of language and syntax-based models use linguistic phrases.

A different approach tries to create hybrid systems that draw on the advantages from rule-based systems and statistical systems to create hybrid machine translation (HMT) systems that produce more accurate translations than either component system.

1.3 Contribution

Hierarchical phrases are encoded in a tree structure just as linguistic trees. Most rule-based translation systems also encode the analysis of a sentence in a tree. While the trees are generated differently, the rules generating hierarchical trees are inferred from unlabeled corpora and RBMT systems use handcrafted rules, alignments between nodes and subtrees on the target side can be computed. Based on the computed alignments, substitution can be performed between the trees.

I propose a post-editing approach based on structural similarity. The tree structures are aligned and subtree substitution based on the similarity of subtrees can be performed. The knowledge-poor approach is compatible with the surface-near nature of SMT systems, does not require other information than what is available in the output and ensures that the approach is generic so it can, in principle, be applied to any language pair.

Three methods of computing the structural alignment are presented.

Lexical substitution is used to improve the translation hypotheses.

Subtree substitution is a graph matching approach where node alignment is computed by comparing leaf nodes and computing a confidence measure based on the similarity between descendants.

TED-R Tree Edit Distance, a well-known distance metric that can be used to compute the similarity between tree structures, is modified to compute the alignment between the root nodes of subtrees. An alignment between nodes can be extracted from an edit script created when computing the tree edit distance, which can be used to guide subtree substitution. The substitution is applied and uses a re-ranking module to search for the best translation hypothesis. To my knowledge, it is an approach not previously applied to post-editing or hybridisation in the field of machine translation.

Chapter 2

Machine translation

This chapter will describe the theory of MT. Section 2.1 describes RBMT and the RBMT component system. The grammatical formalism used to implement the rule-based approach is described in Section 2.1.7. Section 2.2 describes the foundations of statistical MT, state-of-the-art MT and the hierarchical translation model which is used in the SMT component is described in Section 2.2.2.

2.1 Rule-based machine translation

In RBMT, the source language is parsed, the information in the parse tree transferred to a parse tree in the target language and, depending on the linguistic information in the parse tree, a translation is generated on the target side. There is a relation between the complexity of the parsing step and the complexity of the information encoded in the transfer rules. If the source parse includes a large amount of linguistic knowledge, the transfer rules can be less complex. The Vauquois Triangle in Figure 2.1 visualises this relationship and it also shows the 3 phases of the rule-based approach: analysis, transfer and generation phase.

As the source language analysis becomes more complex and uses deep linguistic knowledge, a decreasing amount of transfer knowledge needs to be encoded in the transfer rules. Inversely, if a system only uses direct transfer rules, i.e. transfer rules which apply to words, all linguistic knowledge must be encoded in the transfer rules. If the RBMT system uses interlingua, no transfer rules are necessary and therefore no transfer knowledge, since the source language sentence analysis has created an interpretation from which a target language sentence can be generated directly. It should be noted that a translation system that only relies on direct transfer is not usually classified as RBMT. As the analysis phase becomes more complex, the rules used in the generation phase will also become more complex. All linguistic levels incorporated into the analysis phase must also be incorporated into

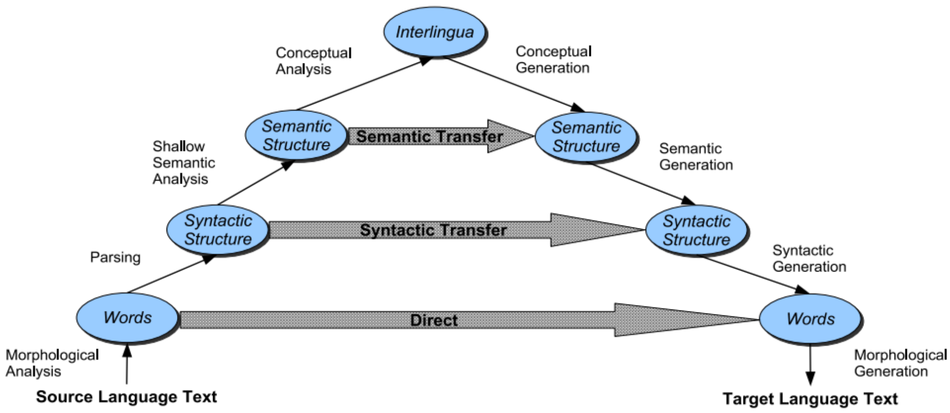

Figure 2.1: Vauquois Triangle

the generation rules.

2.1.1 Morphological rules

Morphological rules extract information about inflection, derivation, lemma, word class, etc. This information is needed to disambiguate words for the application of subsequent rules. Morphological analysis can also be used to extract information about out-of-vocabulary (OOV) words so the RBMT system can produce a translation even though a word is unknown. These rules are not transfer rules as they function solely on the source side.

2.1.2 Lexical categorisation and direct transfer rules

Before firing the transfer rules, multiword expressions are usually identified and parts of speech assigned to words. If the analysis is a step before deeper analysis, word sense disambiguation is attempted if this was not achieved at the morphological level.

If the system uses direct transfer rules, translation is done at word level and corresponds to bilingual dictionary lookup.

2.1.3 Syntactic analysis and transfer

Syntactic analysis will create a syntactic representation of the source language. Syntactic transfer rules work at this level and can here encode reordering, e.g. of adjectives inside a noun phrase in translation from English to French.

The grammar formalism used varies. RBMT systems using dependency, context-free, tree adjoining and head-driven phrase structure grammars have been implemented. At this level, syntactic transformation from source language to target language can also be encoded in the transfer rules. If the source and target languages are from the same family, deeper linguistic analysis may not be necessary to model syntactic transformation.

2.1.4 Semantic analysis and transfer

Shallow semantic analysis such as semantic role-labeling can be used to enrich a parse tree or modify the structure of a tree. A linguistic phrase which fulfills a certain semantic role may also be subject to syntactic transformations or reordering. This can be encoded in the structure and used in the transfer rules.

2.1.5 Interlingua

The interlingua idea is to parse a language and create a language-independent representation of the meaning of the sentence. From the interlingua representation, the target language sentence with the same meaning can be generated. Logic is usually used to represent the meaning of a sentence. In this approach, no transfer rules are required.

RBMT systems relying on interlingua representation are not as common as transfer-based RBMT systems. Semantic analyses are too sensitive to errors in analyses at previous linguistic levels and are not very robust. As a consequence, most commercial RBMT systems are transfer-based, including the RBMT system used in this thesis.

2.1.6 Problems of RBMT

There are often exceptions to many grammatical rules and as human languages are always evolving, more and more exceptions occur which increases the complexity of the grammar.

The development of state-of-the-art rule-based systems take several years and the static nature of the rule-based approach requires a lot of maintenance: new dictionary entries can improve the translation of a sentence, but may introduce errors in other sentences and the rule sets become complex and difficult to maintain. Especially the interaction between rules can become difficult to predict. Lexical items are given new interpretations, new lexical items are introduced into the language while others disappear and as a consequence, the rules must also be updated.

The skill set required to create and maintain rule-based systems is very specific and extensive. The developer must have knowledge of not only linguistics, but software programming and terminology in at least two languages. To create bilingual dictionaries, they must also be able to use lexical information such as parts of speech and morphology. And if a new language pair must be added to the system, transfer-based RBMT systems also face costly and time-consuming development.

```
"<Jeg>"
	"jeg" PERS 1S NOM
	"jeg" N NEU S IDF NOM
"<arbejder>"
	"arbejder" PREF
	"arbejder" N UTR S IDF NOM
	"arbejde" V PR AKT
	"arbejde" N NEU P IDF NOM
"<hjemme>"
	"hjemme" <aloc> ADV LOC
	"hjemme" PREF
"<$.>"
	"." PU @PU
```

Figure 2.2: Cohorts for *I work at home.*

2.1.7 RBMT component system

The rule-based MT system used is GramTrans.[1] The Danish to English translation engine - dan2eng - is described in [Bick, 2007]. GramTrans uses constraint grammar (CG[2]) to analyse the source language and generate the target sentence. It is a grammar formalism which is not easily classifiable as phrase-structure grammar or dependency grammar since it is a methodological approach that does not follow a specific linguistic theory. The rules can be created according to certain linguistic theories, but are not bound to any theory.

Constraint Grammar

CG [Karlsson, 1995] is a rule-based methodology which assigns grammatical tags to tokens and disambiguates based on these tags. The disambiguation can draw on lexical, morphological, syntactic and semantic information depending on the grammatical tags. CG rules are hand-written and handle ambiguity by selecting a *reading* of a token based on the sentence context and rejecting other readings.

The input to a CG parser are so-called *cohorts* which consist of all possible readings of all words. A reading is represented as a list of tags assigned to the word as shown in Figure 2.2 and extracted from a lexicon. An example CG rule would reject the reading of *arbejder* as a noun if the preceding tokens up to the beginning of the sentence and the subsequent tokens do not have a possible reading as a verb. A CG module will apply the CG rules appropriate at this stage of analysis. After rejecting some or all of the wrong readings in the sentence, a different CG module will then assign tags to the

[1] `www.gramtrans.com`

[2] Should not be confused with categorical grammar which has the same abbreviation, but is a very different approach to syntactic description.

readings based on the (partially) disambiguated context. The assigned tags can be functional, structural, semantic, phrasal and other context-sensitive tags. The tagging will again create more readings in the cohort and requires another disambiguation module that takes the new tags into account.

CG modules are applied at every stage and handle tagging, morphology, disambiguation, etc. Because it is a methodology, there are no explicit limitations for what the rules in a CG module can handle. There are also no requirements which tags can be assigned. The tag set is wholly different from implementation to implementation and the final disambiguation procedure also differs.

CG modules can also be embedded. Noun phrases or verb chains, etc. can be identified and a separate CG module applied to perform tagging or disambiguation within that scope before a sentence is passed to the next module.

In using a rule-based approach, CG grammars resemble generative grammars, but CG is a methodological paradigm which more resembles statistical and machine learning systems that do not adhere to any specific linguistic theory. CG also treats syntactic function as primary over syntactic structure and is — as a paradigm — related to dependency grammar.

GramTrans

GramTrans targets unrestricted texts and uses CG to create the representations needed for transfer. Two grammars are used in the Danish to English translation engine dan2eng: DanGram and EngGram. In [Bick, 2007], the CG grammar used (ENGCG) was licensed from LingSoft. EngGram has since replaced ENGCG but the system architecture is the same as described in [Bick, 2007].

The Danish CG parser [Bick, 2001] is a development of Palavras [Bick, 2000] and has inherited some of the rules from the original parser. The tags used by DanGram for disambigutation include tags for lexical category, tense, inflection, gender, functional markers and possibly even phrasal category[3]. Also markers for attachment and dependency are added for syntactic description. The parser consists of (in order of application):

- A morphological analyser: identifies word class, inflection, derivation, composites, fixed expressions.
- Morphological disambiguator: rejects or selects readings proposed by the morphological analyser.

[3] `http://beta.visl.sdu.dk/visl/da/info/tagset_da.pdf`, `http://beta.visl.sdu.dk/visl/da/info/dansymbol.html` and `http://beta.visl.sdu.dk/visl/da/info/dansymbolcg.html`

```
Jeg [jeg] PERS 1S NOM @SUBJ> #1->2
arbejder [arbejde] <mv> V PR AKT @FS-STA #2->0
hjemme [hjemme] <aloc> ADV LOC @<ADVL #3->2
. [.] PU @PU #4->0
```

Figure 2.3: The readings for the sentence in Figure 2.2 after full parse and dependency annotation. The dependency annotation is identified by the #-character.

- Syntactic "mapper": assigns functional and syntactical tags based on the morphological disambiguation.
- Syntactic disambiguator: rejects or selects readings with tags added by the syntactic mapper.
- Dependency annotator: assigns dependency tags to a reading.

The dependency tags represent the head-dependent relationship as shown in Figure 2.3, where the numbers reflect sentence position. Both the subject *jeg* and the adverbial modifier *hjemme* are dependents of the verb *arbejder*. Also note that the verb is the dependent of the root at position 0.

The English CG grammar EngGram [Bick, 2009] works similarly. The input is the readings selected by DanGram, which means the system uses lexical transfer rules. The lexical transfer rules are also encoded in CG but GramTrans works as a black box and the transfer rules are not available.

After transfer, target side lexical movement rules are applied, composites are treated and a morphological generator assigns inflexion. Structural transfer knowledge is implemented as syntactic movement rules that are applied on the target side and syntactic transformation is based on the dependency and syntactic function tags from DanGram.

GramTrans is created to be robust and produce as many dependency markings as possible to be used for disambiguation. Errors in the assignment of functional tags propagate to the dependency level and can result in markings that will produce a dependency tree and a number of unconnected subgraphs with circularities. This presents a problem if the dependency markings are the basis for creating a dependency tree because it is not straight-forward to reattach a subgraph correctly, when the grammatical tags cannot be relied upon.

2.2 Statistical machine translation

SMT uses a translation model and a language model to model the translation process. This combination of translation model and language model is known as the *noisy channel model*. It is extensively used in speech recognition where, instead of a translation model, an acoustic model is combined with a language model. With respect to machine translation, the underlying assumption is that a sentence in the source language was originally a sentence in the target language, but the sentence has been passed through a noisy channel which distorted the target sentence into the source sentence.

To model the distortion in the noisy channel, SMT conditions the probability of a target sentence being a good translation on the source sentence. This leads to the posterior probability of a target sentence given a source sentence $P(t|s)$.

Finding the t that maximises $P(t|s)$ in this expression is difficult but using Bayes rule, the expression can be rewritten as:

$$P(t|s) = \frac{P(s|t)P(t)}{P(s)} \tag{2.1}$$

This expression can be further reduced by leaving out $P(s)$: The source sentence is fixed, so $P(s)$ is a constant and only scales the value of the equation. A target sentence that gives the highest value for the equation will also give the highest value for the numerator. The resulting equation for calculating the best translation is

$$\hat{t} = \operatorname*{argmax}_t P(s|t)P(t) \tag{2.2}$$

$P(s|t)$ and $P(t)$ correspond to the two basic components of a SMT system: translation model and language model, respectively.

The translation model ensures that the correct words or phrases are used in the translation. The translation model usually uses the bag-of-words approach and disregards word ordering and syntax in the target language. The language model ensures a measure of fluency and that the translations are acceptable with respect to word order. During the recombination of translated words, the language model score increases the posterior probability for translations that respect the word order of the target language. A decoder uses the output of both the language and translation models to search for the best translation. In Equation (2.2), the decoder is represented by the maximisation. The decoder searches for the translation that results in the highest probability given language model and translation model scores.

2.2.1 Word-based models

In the beginning of SMT, five models were developed to model the translation process. They are described in [Brown et al., 1993] and are known as the IBM models. The models differ only in the translation model, while the language model stays the same and these models are the foundations of modern SMT. In [Brown et al., 1993], the translation process is modelled by finding all possible word alignments in a sentence:

$$P(s|t) = \sum_{a} P(s,a|t) \tag{2.3}$$

A word alignment a is a correlation or link at word level between the source and target language. An example of a word alignment is shown in Figure 2.4.

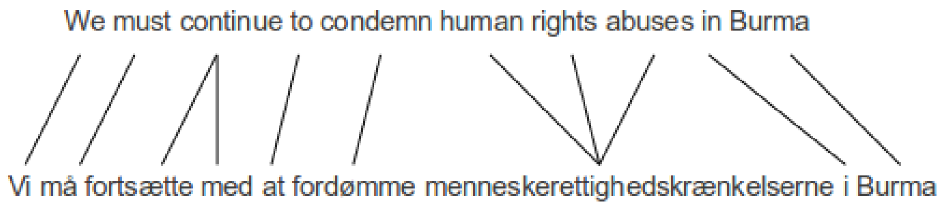

Figure 2.4: An example of alignment between Danish and English

Word alignments can be straightforward 1-to-1 correlations, but often a source language word can be aligned to several words in the target language and vice versa. The occurences of 1-to-many, many-to-1 and many-to-many alignments are more frequent when translating between languages that are not of the same language family e.g. English and Chinese. However, in Danish, like in German, compounding is productive and 1-to-many alignments are common from Danish compounds to complex English noun phrases as can be seen in Figure 2.4. It is also possible to have words that are aligned to an empty or spurious word. This can happen with punctuation and function words which might be necessary in the source language but not in the target language.

The IBM models can be divided into 2 groups by how translation is modelled.

IBM model 1 and 2

In models 1 and 2, the translation is modelled by dividing the probability in Equation (2.3) into 3 different probabilities:

$$P(s,a|t) = P(m|t) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, s_1^{j-1}, m, t) P(s_j|a_1^{j}, s_1^{j-1}, m, t) \tag{2.4}$$

m is the length of the source sentence, s_1^{j-1} are the source words up to position j-1 from the start of the sentence and a_1^{j-1} are the corresponding alignments. The probability from Equation (2.3) is modelled as:

$P(m|t)$ probability of the length of the source sentence given the target sentence

$P(a_j|a_1^{j-1}, s_1^{j-1}, m, t)$ probability of a word alignment given the previous alignment, previous source word, the chosen length of the source sentence and the target sentence

$P(s_j|a_1^j, s_1^{j-1}, m, t)$ probability of a source word given a word alignment and all the previous word alignments, previous source words, the length of s and the target sentence

In words, a length of the source sentence is chosen and then for each position in the source sentence the most probable alignment is determined and a word is chosen to insert at that position.

All sentence lengths are equally probable and $P(m|t)$ is therefore a constant. Model 1 assumes that all word alignments are equally probable as well, so aligning a word in the beginning of a source sentence to the last word in a target sentence is as probable as if they had the same sentence position. Model 2 conditions the alignment probability on the source sentence position and the chosen length of the sentence.

IBM model 3, 4 and 5

For the remaining IBM models, the calculation of the probability from Equation (2.3) has been redefined to include other factors. The simple formulation is:

$$P(s, a|t) = \prod_{i=1}^{l} n(\phi_i|t_i) \prod_{j=1}^{m} t(s_j|t_{a_j}) \prod_{j=1}^{m} d(j|a_j, m, l) \tag{2.5}$$

ϕ is called *fertility*. The fertility of a target word determines how many words it is translated to and $n(\phi|t_i)$ is the probability of a fertility e.g. 1, 2, 3 etc. given a target word at position i in the target sentence. t is the probability that a source word is the translation of a target word. It corresponds to the probability from Equation (2.4). d is the *distortion* probability. It is the probability that a word in the source sentence appears at position j given the position of the translation in the target sentence and the sentence lengths, where l is the length of the target sentence. So a translation is modelled as follows: determine how many words each word is translated into, choose the word forms and determine at which sentence position to insert the word forms.

Models 3, 4 and 5 take into account *spurious words*. Spurious words are words in the target language that are not generated by a source word i.e. they are aligned to `NULL`. There should also be an associated cost for inserting spurious words as well as a cost for permuting spurious words into their target words at their final positions. Finally, when the fertility of a word is larger than 1, some information of the generation is lost because it is unknown whether the generated words were generated in the right order or whether they were generated and then permuted.

Adding terms to Equation (2.5) to model spurious words will yield:

$$P(s,a|t) = \binom{m-\phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \times \frac{1}{\phi_0!} \times \prod_{i=0}^{l} \phi_i! \times \prod_{i=1}^{l} n(\phi_i|e_i) \times \prod_{j=1}^{m} t(s_j|t_{a_j}) \times \prod_{j=1}^{m} d(j|a_j,m,l) \tag{2.6}$$

p_1 and p_0 are the probabilities of generating a spurious word and not generating a spurious word, respectively. ϕ_0 is the number of spurious source words and $m - \phi_0$ is the number of non-spurious words.

The decision to generate spurious words must be made for all positions in the source sentence after the first word and leads to the first new term in Equation (2.6). The second new term is the cost of inserting the spurious words and the third new term multiplies the rest of the equation by how many different ways the 1-to-many translation could be generated according to the model.

IBM model 3 models the distortion probabilities based on absolute positions. Model 4 conditions on relative distortion and can model tendencies such as movement in translation, e.g. that a translation has a tendency to move to the left when translating a given language pair. The word alignment is also conditioned on the previous alignment in model 4 while model 3 only uses sentence positions.

Models 3 and 4 are *deficient*. This means that probability mass is wasted on impossible events such as translations were all source words are aligned to one target word. In model 4, the distortion modelling does not take sentence boundaries into account and will waste probability on movement out of the sentence. Model 5 tries to mend this.

2.2.2 Phrase-based MT

There are a few issues with the word-based models: the alignments produced by the IBM models are asymmetrical — a source word can be aligned to one target word only, i.e. many-to-1 alignments are possible, while 1-to-many and many-to-many alignments are not. This means word-based models cannot model the alignment in Figure 2.4.

All IBM models use words as the basic unit of translation. However, the biggest advancement in SMT has been to use a *phrase* as the basic unit of translation [Koehn et al., 2003]. Note that a phrase in a SMT context is not a linguistic phrase like NP, VP, etc. but a string of continuous words also known as an *ngram*. Ngrams are of differing lengths e.g from one word(unigram) to five words or more. This fact makes modelling 1-to-many and many-to-1 alignments easier. Local reordering is also more easily modelled using ngrams than distortion probability. Yet another improvement is the modelling of spurious words. The IBM models handle spurious words by aligning some words with `NULL` and to add words in the target sentence. Since alignment in phrase-based SMT holds across ngrams of different lengths, this is no longer necessary.

To find phrase pairs, word-based alignments are used. A bidirectional word alignment is computed e.g. with IBM model 2 and phrase pairs are found using these alignments. A source phrase and a target phrase constitute a phrase pair iff the words in the phrases are aligned to each other and not to words outside of either phrase.

To estimate the phrase translation probability, the relative frequency of a phrase pair is calculated.

$$\phi(s_i|t_i) = \frac{count(s_i|t_i)}{\sum_{s_i} count(s_i|t_i)} \tag{2.7}$$

The subscript in s_i denotes a phrase from a source sentence S and similar for t_i.

During translation, a source language input sentence is divided into a sequence I of phrases s_i^I and each source language phrase is translated into a target language phrase t_i^I. $P(s|t)$ from Equation (2.2) is decomposed into

$$p(s_1^I|t_1^I) = \prod_{i-1}^{I} \phi(s_i|t_i) d(a_i - b_{i-1}) \tag{2.8}$$

$\phi(s_i|t_i)$ is the translation probability from Equation (2.7), a_i is the start position of the current source phrase, b_{i-1} is the end position of the source phrase that was translated into the $i - 1th$ target phrase and $d(a_i - b_{i-1})$ models the distortion probability.

SMT has often been criticised for the lack of linguistic knowledge incorporated into the translation modelling. It seems that some phenomena in language require linguistic knowledge, e.g. global reordering. Integrating linguistic knowledge into SMT is not without problems due to the surface-near approach of SMT, but some approaches show promising results for introducing additional linguistic information at the word level and syntactic structure: Factored translation models and hierarchical phrases.

Factored SMT

Factored translation models [Koehn and Hoang, 2007] use vectors as the basic unit of translation, not words or phrases. Each vector represents a word tagged with word-level information such as morphology, part-of-speech

$$X_1 \text{ i øvrigt } X_2 \longrightarrow \text{moreover, } X_1\ X_2$$

Figure 2.5: A rewrite rule learned from training on the Europarl corpus.

etc. Factored SMT has shown improved translation with sparse data, but the improvement wears off as the amount of data is increased. A significant improvement has been seen when translating case-unspecified English to German for noun phrase agreement errors.

Though all the linguistic information is added to a token, the decoding is still phrase-based.

The vector representation of words is similar to the readings generated from EngGram and DanGram. An interesting approach could be to use a CG parser to tag training data and thus be able to incorporate linguistic information above word level into the vector representations. This would however follow the intuition in [Chen and Eisele, 2010] where the rule-based system adds hypotheses to the phrase-table in an attempt to reverse-engineer the rule-based system. This is not possible according to the research license for GramTrans which prohibits reverse-engineering the rule-based system. With a different CG parser, this would still be feasible.

Hierarchical phrase-based SMT

Hierarchical phrases are phrases that can contain subphrases, i.e. a hierarchical phrase contains non-terminal symbols. An example rule from Danish to English is shown in Figure 2.5.

X_n is a nonterminal and the subscript identifies how the nonterminals are aligned. While this particular phrase is linguistically motivated and could look like a linguistic rule, many hierarchical phrases learned from bitexts do not.

This type of reordering is interesting because using traditional phrase-based SMT, there is little improvement in using phrases longer than trigrams [Koehn et al., 2003]. The fact that performance improves when using phrases up to trigrams, stagnates over phrases up to six tokens in length and decreases over longer phrases indicates data sparsity as a source problem. It is often desirable to model the reordering in a sentence for more than a few consecutive words.

As mentioned, the hierarchical phrases learned are not linguistic. They are learned from bitext with unannotated data. Phrases in hierarchical models are formally productions from a synchronous context-free grammar (SCFG) and can be viewed as a move towards syntax-based SMT [Chiang, 2005]. Since hierarchical phrases are not linguistic, Chiang makes a distinction between *linguistically* syntax-based MT and *formally* syntax-based MT where hierarchical models fall in the latter category because the structures they are defined over are not linguistically informed, i.e. unannotated bitexts.

A hierarchical model is based on a SCFG and the elementary structures are rewrite rules:

$$X \longrightarrow \langle \gamma, \alpha, \sim \rangle \tag{2.9}$$

As above, X is a nonterminal, γ and α are both strings of terminals and nonterminals and $\sim$ is a 1-to-1 correspondence between nonterminals in γ and α. As in Figure 2.5, the convention is to use subscripts to represent $\sim$.

To maintain the advantage of the phrase-based approach, *glue rules* are introduced to the rules that are otherwise learned from raw data.

$$S \longrightarrow \langle S_1 X_2, S_1 X_2 \rangle \tag{2.10}$$

$$S \longrightarrow \langle X_1, X_1 \rangle \tag{2.11}$$

As the only rewrite rules, they contain the nonterminal S. These rules are added to give the model the option of combining partial hypotheses serially and they make the hierarchical model as robust as the traditional phrase-based approaches.

To train the model, *initial phrase pairs* must be identified from which the phrase translation parameters can be estimated. The criteria for finding phrase pairs in phrase-based SMT applies for finding initial phrase pairs:

1. $s_k \sim t_{k\prime}$ for a $k \in [i, j]$ and $k\prime \in [i\prime, j\prime]$
2. $s_k \nsim t_{k\prime}$ for a $k \in [i, j]$ and $k\prime \notin [i\prime, j\prime]$
3. $s_k \nsim t_{k\prime}$ for a $k \notin [i, j]$ and $k\prime \in [i\prime, j\prime]$

For a word aligned sentence pair $\langle s, t, \sim \rangle$, a rule $\langle s_i^j, t_{i\prime}^{j\prime} \rangle$ is an initial phrase pair if the above definition holds.

In words: a word inside a phrase in a candidate initial phrase pair must be aligned to a word in the aligned phrase (1). If it is the case that a word in the source phrase is aligned to a target word outside the aligned target phrase (2) or vice versa (3), the candidate is not an initial phrase pair.

To generate all different phrase pairs, we define the set of rules in $\langle s, t, \sim \rangle$ as the smallest set satisfying:

1. If $\langle s_i^j, t_{i\prime}^{j\prime} \rangle$ is an initial phrase pair, then $X \longrightarrow \langle s_i^j, t_{i\prime}^{j\prime} \rangle$ is a rule
2. If $r = X \longrightarrow \langle \gamma, \alpha \rangle$ is a rule and $\langle s_i^j, t_{i\prime}^{j\prime} \rangle$ is an initial phrase pair so $\gamma = \gamma_1 s_i^j \gamma_2$ and $\alpha = \alpha_1 t_{i\prime}^{j\prime} \alpha_2$ then $X \longrightarrow \langle \gamma_1 X_k \gamma_2, \alpha_1 X_k \alpha_2 \rangle$ is a rule and k is not used in r

This creates too many rules which slow down decoding and results in spurious ambiguity - distinct derivations result in the same translation and similar model feature vectors. The grammar will be filtered according to the following heuristics:

1. If there are initial phrase pairs with the same alignment points, only the smallest is kept.
2. Initial phrase pairs are restricted to length 10 on the source side and a rule to five terminals/nonterminals on the source right-hand side.
3. In (1), s_i^j must be length > 1.
4. Rules can have at most two nonterminals. This reduces the complexity of the decoding.
5. Rules must always have some aligned words. This ensures that translation decisions are based on lexical evidence.

Decoding will be done with a parser and a post processor that maps source language derivations to target language derivations and beam search to prune the search space.

2.2.3 Problems of SMT

To build good systems, large amounts of bilingual and monolingual data are necessary. But bitexts with a large number of sentence pairs are not available for all languages. Noise in the bitexts such as disfluencies and parallelisation errors can result in wrong alignments and the source of the errors can be difficult to track down.

Language models can produce fluent output, while the translations can be incorrect because the language model works with monolingual data. The training, tuning and decoding in SMT systems require resources that are not usually readily available and if not, the training and tuning may become time-consuming.

And lastly, there is very little linguistic knowledge in SMT. Factored models and syntax-based SMT tries to incorporate linguistic knowledge, but hierarchical phrases may also not be linguistically motivated while linguistically syntax-based decoding become too restrictive and parsing restraints must be relaxed or otherwise the robustness of SMT is lost.

Chapter 3

Structural similarity

The post-editing approach relies on structures output by the component systems. It is necessary to find similar structures to perform subtree substitution. Matching structures is a problem in several application areas such as semantic web, schema and ontology integration, query mediation etc. Structures include database schemas, directories, diagrams and graphs. [Shvaiko and Euzenat, 2005] contains a comprehensive survey of matching techniques. A commonality between hierarchical trees and database schemata is the lack of explicit semantics. There is no information in the trees except surface forms, phrase alignment and unlabeled relations. The dependency trees contain much richer information. In addition to the information in the hierarchical trees, dependency trees include labelled dependencies and, in the case of the dependency trees from GramTrans, syntactic function, semantic primitives and parts of speech for gender, lexical and phrasal class. In combination, the tags for syntactic function and dependency provide labelled dependencies.

The matching operation determines an alignment between two structures and an alignment is a set of *matching elements*. A matching element is a quintuple: $\langle id, e, e', n, R\rangle$:

id Unique id.

e, e' Elements from different structures.

n Confidence measure.

R The relation holding between the elements.

External resources, information contained in the structures or pre-existing alignments can be used in the matching process. The resources that can be used in the matching process are shown in Figure 3.1.

o and o' are the structures to be matched, A is an optional existing alignment, r is external resources, p is parameters, weights and thresholds and A' is the set of matching elements created by the process.

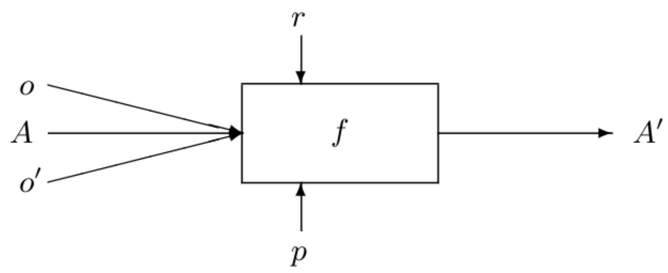

Figure 3.1: The matching process.

When computing alignment in this thesis, only matching elements with an equivalence relation (=) are used. Some systems can return subsumption relations or incompatibility relations, but they would require more information than is present in the hierarchical trees.

Thresholds are not used in the thesis. Each candidate will return a match and if no exact match can be found, string matching is used to create a confidence measure for the candidate element and the best matching element is then returned.

The returned alignment can be a new alignment or a refinement of A. o will be a dependency tree and o' the hierachical trees from the SMT component system. The phrase alignment between source and target language is used to build the initial alignment A by linking the respective nodes in dependency and hierarchical trees. This is however not always a very refined alignment. Figure 3.2 shows an alignment between a dependency tree and a hierarchical tree.

The matching elements are correct as far as surface forms can indicate when looking at nodes in isolation. Only leaf nodes from the hierarchical tree are included in the alignment. To match phrases or subtrees, an alignment must be found between nonterminal nodes in the trees. Therefore, the phrase alignment will function as an initial alignment to guide the matching process. According to the classification in [Shvaiko and Euzenat, 2005], phrase alignment corresponds to a terminological and element-level matching technique. More precise classification is dubious as both string-based methods, tokenisation and word alignment are used.

Shvaiko and Euzenat include in their survey four graph-based techniques: graph matching, leaves, children and relations.

3.1 Graph matching and leaves

Matching graphs can be viewed as an optimisation problem. The task is to find the graph that minimises a distance metric for similarity between two graphs.

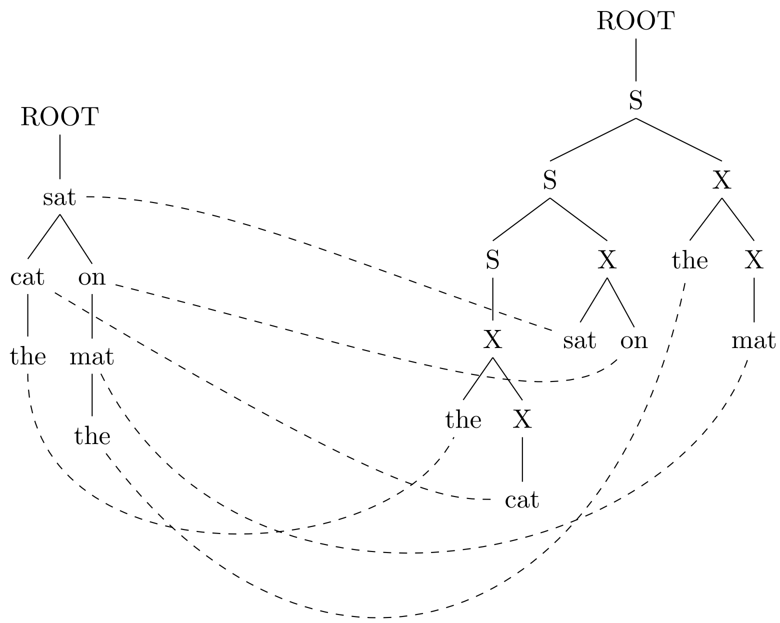

Figure 3.2: The initial alignment from an hierarchical tree (right) to a dependency tree (left).

A well-known algorithm for computing similarity between trees is the *Tree Edit Distance* algorithm, which computes how many operations are necessary for transforming one tree into the other. Following [Zhang and Shasha, 1989] and [Bille, 2005], the operations are defined on nodes and the trees are ordered, labelled trees.

There are 3 different edit operations:

relabel Change the label of a node in a tree.

delete Remove a node n from a tree. Insert the children of n as children of the parent of n so the sequence of children are preserved. The deleted node may not be the root node.

insert Insert a node as the child of a node n in a tree. A specified subsequence of children of n are inserted as children of the new node so the sequence of children are preserved. An insertion is the inverse operation of a deletion.

A cost function is defined for each operation. The goal is to find the sequence of edit operations that turns a tree T_1 into another tree T_2 with minimum cost. The sequence of edit operations is called an *edit script* and the cost of the optimal edit script is the tree edit distance.

The cost functions should return a distance metric. A distance metric must satisfy the following conditions:

1. $\gamma(i \to j) \geq 0$ and $\gamma(i \to i) = 0$
2. $\gamma(i \to j) = \gamma(j \to i)$
3. $\gamma(i \to k) \leq \gamma(i \to j) + \gamma(j \to k)$

The *edit distance mapping* is a representation of an edit script. A rename operation is represented as $(i_1 \to j_2)$ where the subscript denotes that the nodes i and j belong to different trees. $(i_1 \to \epsilon)$ represents a deletion and $(\epsilon \to j_2)$ an insertion.

The cost of an edit distance mapping is given by:

$$\gamma(M) = \sum_{(i,j)\in M} \gamma(i \to j) + \sum_{i \in T_1} \gamma(i \to \epsilon) + \sum_{j \in T_2} \gamma(\epsilon \to j) \tag{3.1}$$

γ is the cost for an edit operation or mapping and $j \in T_2$ means j is in the set of nodes in T_2.

It is important to note that the trees are ordered trees. The unordered version of the tree edit distance problem is NP-hard, while polynomial algorithms based on dynamic programming exist for ordered trees.

The algorithm does not require an input alignment or external resources. The cost functions for deletion, insertion and renaming must be defined on the information present in the nodes and a postorder id must be assigned to the nodes. This id is assigned by traversing the tree depth-first and assigning an integer as id. The algorithm visits each node in the trees in post order and determines based on the cost assigned by the cost functions, which edit operation should be performed.

The tree edit distance is a special case of the *forest distance* where each forest is a single tree. However, when computing the tree edit distance, the distance between the subtrees of all nodes in the dependency tree and hierarchical trees must be computed. The algorithm reduces the computation of the tree edit distance to subproblems where the forest distance are computed.

Following the paradigm of dynamic programming, intermediate computations of the forest distance are stored in an array so each problem only needs to be computed once. Traversing the trees in post order ensures that, when the forest distance between $T_1[i]$ and $T_2[j]$ must be calculated, the forest distance of the subproblems in $T_1[i]$ and $T_2[j]$ are stored in an array and can be retrieved rather than recomputed. The subproblems are rooted in *keyroots*. A keyroot is the root node or a node that has a left sibling and the number of keyroots is equal to the number of leaf nodes in a tree.

A different technique to compute the structural similarity of nonterminal nodes — called *leaves* — is also possible. The structural similarity of the nodes in a matching element is conditioned on the similarity of the descendants of the nodes that are leaf nodes. Because hierarchical leaf nodes have element-level similarities to nodes in the dependency trees, this alignment can be used to guide the computation of structural similarities of nonterminal nodes. For this approach, all dependency nodes are treated as leaf nodes.

3.2 Children and Relations

The structural similarity between two nodes can also be computed based on how similar their children nodes are. With this technique, only immediate descendants are considered. This technique is not appropriate for computing similarity between hierarchical trees and dependency trees. A nonterminal node in a hierarchical tree that has nonterminal nodes as children would not create matching elements of high quality. Nonterminal nodes in hierarchical trees do not have element-level similarities to nodes in dependency trees.

Relations are also not an appropriate basis for computing alignment because only dependencies are typed and similar information is not present in hierarchical trees.

Chapter 4

Hybrid system

This chapter describes existing hybridisation approaches and the hybrid system used in the experiments. Section 4.1 describes how MT systems can be combined, Section 4.2 shortly describes how the RBMT component is used, Section 4.3 describes the SMT component, training, configuration and the corpus used. Section 4.4 describes the matching approaches and Section 4.5 introduces re-ranking and the re-ranking module used.

4.1 Hybrid machine translation

Hybrid machine translation (HMT) is a paradigm that seeks to combine the strengths of data-driven machine translation and knowledge-driven machine translation. The different approaches have complementary strengths and weaknesses [Thurmair, 2009] which have led to the emergence of HMT as a subfield in machine translation research.

The strength of SMT is robustness - i.e. it will always produce an output - and fluency due to the use of language models. A weakness of SMT is the lack of linguistic knowledge, which make cases requiring such information, e.g. long-distance dependencies, difficult to handle.

RBMT systems translate more accurately in cases without parse failure, since they can take more information into account e.g. morphological, syntactic or semantic information, where SMT only uses surface forms. RBMT suffers from a lack of robustness when parsing fails and in lexical selection in transfer. RBMT systems are also very costly to build, and maintenance and development can be very complex due to the interdependency of rules.

MT systems can be combined serially or in parallel. Serial coupling is exemplified with statistical post-editing (SPE). A RBMT engine produces translations and they are edited by a statistical module which is trained on bitexts. MT systems can also be coupled in parallel. A simple approach is to have a number of MT systems produce translations for a n-best list and use a re-ranking module to rescore the translations. Using this approach, the best improvements are achieved with a large number of systems running in parallel and this is not feasible in a practical application, mostly due to the computational resources required by the component systems. The translations will also not be better than the one produced by the best component system. Other approaches try to create better translations by combining translation hypotheses according to different criteria.

Tighter integration of rule-based and statistical approaches have also been proposed: Adding probabilities to parse trees, pre-translation word reordering, enriching the phrase table with output phrases from a rule-based system [Eisele et al., 2008], creating training data from RBMT systems etc. The factored translation models also present a way to integrate rule-based parsing systems.

The architecture of the hybrid system used in this thesis is parallel coupling with post-editing. A diagram of the hybrid system can be seen in Figure 4.1. *Post-editing* is a process conducted by translators and can be used both with translation memories (TM) and with MT systems. A TM system stores previously translated words, phrases and sentences with their source phrase and if a source phrase is encountered again, candidate translations are presented to the post-editor. The post-editor will choose the appropriate translations and translate the remaining untranslated text. When using a MT system, the translation is presented to the post-editor. In the remainder of this thesis, post-editing will refer to automatic post-editing of MT output where the MT output can be e.g. translations, derivations, trees.

Coupling the component MT systems serially or in parallel is more appropriate than using an architecture which relies on integration when one of the systems is a black box. Having access to the target language analysis from the RBMT system means that it is possible to create a tighter coupling than just using the surface forms from the output and opens the possibility for manipulating structures rather than surface forms.

The trees created by Moses and GramTrans are both graph structures. Matching a dependency tree to hierarchical trees or using subgraphs from one structure to improve the other can be used to create better translations. When it is possible to get structures as output from both components, it becomes possible to use algorithms for similarity and alignment between structures. If this alignment can be computed, performing substitutions, relabeling, and other transformations become possible. The quality of these manipulations are conditioned not only on the alignment from one structure to the other, but also on the alignment between source and target in the structures themselves. The nature of RBMT suggests that the alignment between source and target are of high quality, while the alignment between source and target in the SMT output is of varying quality depending on the amount of training data in general and the occurence of words and collocations in the training data.

The post-editing approach proposed here does not exactly fit the classi-

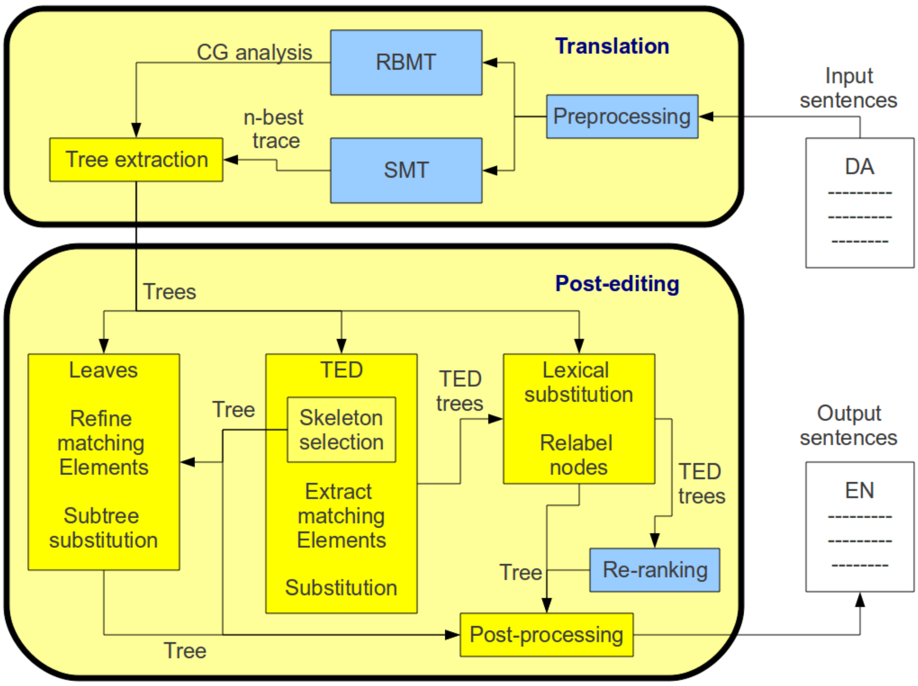

Figure 4.1: Hybrid system architecture.

fication of parallel coupling approaches in [Thurmair, 2009]. Other coupling architectures with post-editing work on words or phrases and generate confusion networks, while the units focused on here are graphs, i.e. tree structures. The approach does, however, select a tree to function as a skeleton upon which transformations are conducted.

4.2 RBMT component

The Danish to English translation engine in GramTrans - dan2eng - is called through an API with HTTP POST and works as a black box. The output from dan2eng is a Constrant Grammar analysis on the target language side after all transfer and target side transformation rules have been applied. Sample output is shown in Figure 4.2. In the analysis, dependency information is provided and they form the basis for creating the tree used for structural similarity computation. Part-of-speech tags, source and target surface structure, sentence position and dependency information are extracted from the CG analysis.

```
<texts>
  <text id="6" sourcelang="dan" targetlang="eng" raw="1">
1  'Vi' "vi" <n1> <left> <1> <*> <H> PERS 1P NOM 'ni' :we @SUBJ  #1->2  [we] ... We
2  'er' "vre" <n2> <2> <va+LOC> <for^vUp-behov> <for^vUp-brug> <for^vkp-udtryk> <vk> <mv>
   V PR 1P AKT 'estas' :are @FS-STA  #2->0  [be] ... are
3  'mange' "mange" <n3> <right> <3> <quant> DET nG P NOM 'multaj' :many @SC  #3->2  [many] ... many
4  $, $,  #4->3
5  'som' "som" <n5> <left> <5> <clb> <rel> INDP P 'kiu' :that @SUBJ  #5->7  [that] ... that
6  'gerne' "gerne" <n6> <left> <6> <amod> ADV 'volonte' :would=like=to @ADVL  #6->7
   [would=like=to] ... would=like=to
7  'ser' "se" <n7> <7> <vq> <+interr> <fra^vp-bort> <som^vta> <mv> <np-close>
   V INF 3P AKT 'vidas' :see @FS-N<  #7->3  [see] ... see
8  'en' "en" <n8> <8> ART UTR S IDF '*' :a @>N  #8->9  [a] ... a
9  'fderation' "fderation" <n9> <right> <9> <HH> <Lciv> <+n> <+af> <H> <neut> N UTR S IDF NOM
   'federacio' :federation @ACC  #9->7  [federation] ... federation
10 'af' "af" <n10> <10> <np-close> PRP 'de' :of @N<  #10->9  [of] ... of
11 'nationalstater' "nationalstat" <n11> <11> <SUF:stat> <Lh> <+n> <for+> <ADJ:national+stat>
   N UTR P IDF NOM 'nacia=sxtatoj' :nation=states @P<  #11->10  [nation=state] ... nation=states
12 $. $.  #12->0
  </text>
</texts>
```

Figure 4.2: Sample RBMT output.

4.3 SMT component

Moses [Koehn et al., 2007] is used as the SMT component. Moses has implemented a CKY+ algorithm for hierarchical decoding of tree-based models, which is the umbrella term for all the implemented models that use formal syntax in the sense of [Chiang, 2005]. Moses can also train models that use linguistic syntax in either target or source language or both.

Flags can be set to change the output from 1-best hypothesis to a n-best list. It was however not possible to get the trace information for the n-best list, but only for the best hypothesis. The trace information contains the derivations which produce the translation hypotheses. The source code for Moses was modified to be able to output trace information from which the n-best hierarchical trees can be reconstructed.

Version 6 of Europarl [Koehn, 2005] was used for training and test data (See subsection 4.3.1). The sentence aligned Danish-English part was used for training, and to tune parameters with MERT, the 2006 test set from the NAACL WMT 2006 was used [Koehn and Monz, 2006]. GIZA++ compiled hierarchical phrases which were extracted by Moses to train a translation model and SRILM [Stolcke, 2002] was used to train the language model component. The training of the entire SMT system was done with the Experimental Management System [Koehn, 2010] distributed with Moses. The configuration for the training of Moses followed the standard guidelines in the syntax tutorial.[1]

To train the SRILM language model, the English side of Europarl was

[1] `http://www.statmt.org/moses/?n=Moses.SyntaxTutorial`

Table 4.1: Corpus statistics.

Aligned sentences	1,785,775
Danish words	46,102,455
English words	48,833,481

used.

4.3.1 Europarl

Europarl contains plenary speeches extracted from the proceedings of the European Parliament and includes 21 European languages as of version 6. Danish is a low-resource language when it comes to language resources, especially parallel corpora. As a consequence, the best option is to use Europarl even though it is domain specific. For training, the prepared version of Europarl was used. The corpus statistics can be seen in Table 4.1.

The 4th quarter of 2000 has been left out in the prepared version. From this subcorpus, several test sets have been extracted which can be used for tuning and testing.

4.4 Matching approach

An important decision regarding this type of hybridisation is how to compute the alignment and the size of the substituted subtrees. Another important decision is what criteria to use for substitution and how to constrain the substitution. Irrespective of which technique is chosen to compute structural similarity, the resulting alignment should be refined to contain matching elements to nonterminal nodes as shown in Figure 4.3.

There are potential sources of errors at all steps in this setup. Beyond the simple sources such as either of the components fail to produce an output e.g. because of downtime, unstable internet connection and other resource requirements, alignment is a main source of problems.

As mentioned earlier, the dependency markings in the RBMT output do not always make up only one tree. Propagation of errors in the functional tags means that separate graphs with circularities are created in addition to a dependency tree. For a dependency tree that spans the entire sentence, it must be possible to traverse the tree from all leaf nodes up to the root node. If it is not possible, the node which has been traversed previously is identified. The dependency is changed to the root node such that the word order is respected.

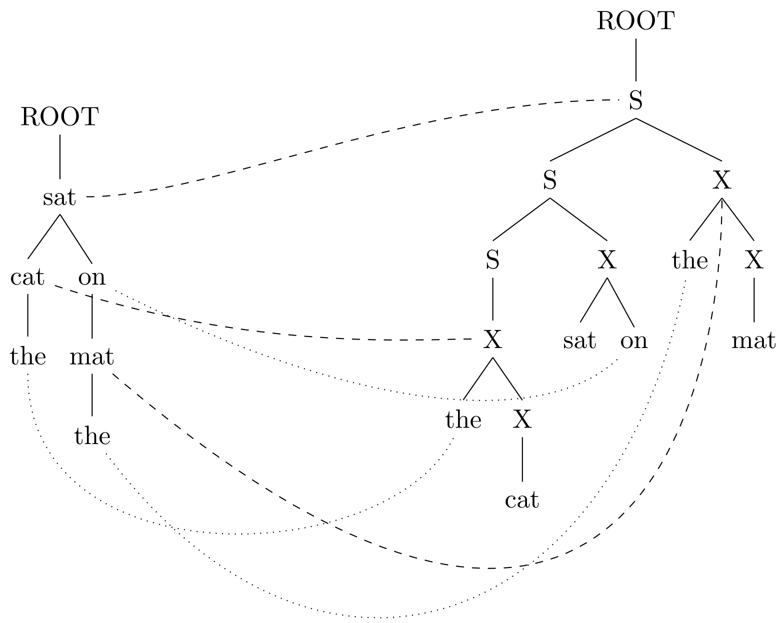

Figure 4.3: The refined alignment from dependency tree to hierarchical tree.

```
offentliggoeres X : X -> will be published X : 1-3
```

Figure 4.4: Simplified example of a simple alignment.

4.4.1 Alignment challenges

The Moses decoder can output word-to-word alignment because phrase alignment is based on it. However, the changes made to the chart decoder to output the n-best trace information is simple and does not output the alignment information. Currently, the tree extraction module computes an alignment between the source and target language phrases.

Alignment problems occur early in the pipeline. The segmentation of words into phrases done by Moses does not always correspond to the word-based segmentation required by the CG parser, simplex phrases recognised by the CG parser rarely correspond to phrases in Moses and the hierarchical phrase alignment is not easy to handle.

Aligning hierarchical phrases like the one in Figure 4.4 is not complicated. The ordering is identical and the Danish word *offentliggøres* is aligned to *will be published*. The numbers `1-3` refer to the alignment of non-terminal nodes based on phrase positions.

A more complicated alignment problem is shown in Figure 4.5. While it is straightforward to decide the alignment of surface forms by visual inspection, handling the alignment computationally is not. There are two methods of handling this type of alignment appropriate for the component systems. Because there are an equal number of tokens in the English phrase and Danish phrase, aligning the tokens 1-1 without reordering would be a solution that, in this case, results in a correct alignment. This has the advantage of corresponding to the word-based CG approach in GramTrans.

There is not a high degree of reordering between Danish and English because they are both SVO languages and closely related Germanic languages, but Danish is a verb-second language. The finite verb in a Danish declarative sentence must be the second constituent. This is not the case in English. Another approach relies on weak reordering and would align *findes* with *there are*. This reduces the alignment problem to aligning *vi der* with *we*. In this case, the alignment is noisy, but usable matching elements are produced. Both approaches are implemented in the hybrid system and the first approach supercedes the second due to the advantage of correlating with the CG approach.

```
vi X der X findes : X -> X, we X there are : 1-3 3-0
```

Figure 4.5: Example of complex alignment.

The noise problem propagates to the matching elements based on the source to target language alignments and, as a consequence, has an impact on the computation of the refined matching elements.

The matching elements that make up the initial alignment are all computed based on the source phrase. Additional information from the dependency trees could be used in the matching approach. In this system, the amount of information used is restricted to the information readily available in the trees. When using source language phrases to create the initial alignment, i.e. the alignment between the dependency tree and hierachical trees over the source phrases, there are three different cases to take into account:

1. The source phrase in node i is identical to the source phrase in node j.

2. The source phrase in i is contained in the source phrase in j.

3. The source phrase in i contains the source phrase in j.

One could argue that cases where some of the source words are unaligned in both i and j should be handled. Here, we take the approach as described when finding initial phrase pairs in hierarchical SMT and discard such candidates.

In case (1), matching elements with source phrase identity is assigned a confidence value of 1.0. In case (2) and (3), the unigram overlap is computed and normalised over the length of the longest source phrase to give a measure of the overlap. All other candidates are discarded.

The sentences in Europarl can be long and more than one candidate matching element can be produced for a single node in an hierarchical tree. In these cases, the candidate matching element with the smallest difference in sentence position is chosen. This approach is only possible if there is weak word reordering between source and target language.

Future work should attempt improve the modification to the Moses chart decoder to output the word alignment. The refined alignment relies on the initial alignment, which in turn relies on the alignment between source and target language. Therefore, improving the source to target language alignment will also improve the performance of the post-editing approach.

4.4.2 Alignment refinement

Not all matching elements in the initial alignment can be refined. If a leaf node in the dependency tree is aligned to a leaf node in a hierarchical tree, no refinement is needed. Criteria for selecting initial matching elements for refinement are needed.

In the RBMT output, there are no indications of where the parser encountered problems. If a surface form is an OOV word, the morphological analyser is used to assign a lexcial category based on the word form, hypothesise additional tags based on the analysis and proceed with parsing. There are no tags or confidence measures to indicate these problems.

In the SMT output, a marker is appended to a surface form to indicate that the word has not been translated. This is necessary to distinguish cases where the surface form is identical in both source and target language and when translating names. The marker gives an indication of where enriching a hierarchical tree with RBMT output can result in improvement of translation quality.

This does not handle cases where the SMT component system chose a wrong translation option. Handling cases where the wrong translation hypothesis has been chosen is not easy in the knowledge-poor approach that has been adopted in the system. Part-of-speech tags or confidence measure tags can give an indication, but neither are available in both trees. Again, only in the hierarchical tree is it possible to find translation hypotheses where the score assigned by the SMT decoder is low. It is also unknown which component system has chosen the more accurate translation, i.e. it is possible that the phrase represented by the dependency subtree is a less accurate translation than the hierarchical subtree.

Listing 4.1: Pseudo-code for computing a refined alignment based on the leaves approach.

```
for untranslated, dep in match_elements:
  if untranslated not in hier_tree:
      continue
  dep_tree = subtree(dep)
  dep_tree = isolate_continuous_tree(dep_tree)
  hier_subtree = tree_alignment(dep_tree, hier_tree)
  if leaf_node(hier_subtree) == True:
    hier_tree.relabel_node(hier_subtree, dep)
  else:
    insert_position = hier_subtree.root.pos
    hier_tree.insert_node(dep, insert_position)
    hier_tree.remove(hier_subtree.root)
```

4.4.3 Subtree substitution

Based on the observations above, hierarchical trees are chosen to function as skeletons. Substituting dependency subtrees into a hierarchical tree is more straightforward than using dependency trees as skeletons. Removing subtrees from hierarchical trees and inserting dependency subtrees does not destroy the linguistic information in the tree. It is even possible to transform the subtree, based on the sentence positions in the subtree, into a hierarchical style tree with a dummy nonterminal node as parent for all nodes in the dependency subtree.

Leaves

Based on the OOV marker, the matching technique based on leaf nodes is implemented to refine matching elements between dependency and hierarchical trees and based on this alignment, substitute the hierarchical subtree with a dependency subtree. Pseudo-code for the substitution algorithm can be seen in Listing 4.1.

The `if`-statement is necessary due to the fact that a previous substitution may remove an untranslated node whose matching element has not yet been iterated over. The dependency subtree is identified by collecting all descendants. The descendants are handled as leaf nodes because both leaf and nonterminal nodes contain surface forms in a dependecy tree.

The dependency trees provided by GramTrans are not always projective. A continuous subtree must be isolated before an alignment between subtrees can be found because the hierarchical trees resemble phrase structure trees in the sense that discontinuous phrases are not allowed. Discontinuous phrases are handled using glue rules.

To identify the corresponding subtree in the hierarchical tree, the matching elements that contain the nodes in the dependency subtree are collected and a path from each leaf node to the root node is found. The intersection of nodes is retrieved and the root node of the subtree identified. It is not always possible to find a common root node besides the root node of the entire tree. To prevent the loss of a high amount of structural information, the root node cannot be replaced or deleted. This technique is only used in this approach due to the fact that naively replacing the identified root node of the hierarchical subtree can remove too many nodes and lead to a severe loss of translation accuracy.

In cases where two leaf nodes are still aligned after refinement, a rename operation is carried out on the hierarchical node instead of a substitution. When the two subtrees are not leaf nodes, the dependency subtree is inserted before the hierarchical subtree — to ensure that the subtree is inserted at the right position — and the hierarchical subtree is subsequently removed.

It seems that using the leaves technique is appropriate when the structures are diverse as is the case of hierarchical trees and dependency trees. If the structures were less diverse, the techniques mentioned in Section 3.2 might be more appropriate, e.g. in the case of matching dependency trees.

4.4.4 Substitution based on an edit script

An edit script is a representation of the tree edit distance. One of the edit operations is renaming. The set of rename operations in an edit script can be used as matching elements.

The cost functions for the edit operations are modified to assign a lower cost to rename operations where one of the nodes is a nonterminal node from a hierarchical tree. This violates the restrictions which ensure that the edit distance is in fact a distance metric. The cost function for the rename operation can be seen in Listing 4.2.

To generate matching elements that align dependency nodes to nonterminal hierarchical nodes, the renaming cost is low for renaming nonterminals from hierarchical trees. If the two nodes have the same target and source phrase, the rename operation does not incur any cost and neither does the renaming of untranslated phrases. While this does not favour alignment to nonterminal nodes, it ensures that the matching elements from the initial alignment that does not require refinement are not altered. Also, if the source is the same and the difference in sentence position is no more than five, the renaming cost is reduced. Experiments showed that a window of five words was necessary to account for the difference in sentence position incurred by reordering and difference in segmentation by the component systems. At the same time, the window size ensured that nodes which should be aligned with a different node with the same source phrase were not assigned a lower cost. If the nodes fullfil these requirements, the matching elements will be of higher quality.

Listing 4.2: Pseudo-code for computing renaming cost.

```
def rename(dep, hier):
  if nonterminal(hier) == True:
    return 0.5
  elif untranslated(hier.tgt) == True:
    return 0
  if hier.src == dep.src and diff(hier.pos,dep.pos) < 6:
    if hier.tgt == dep.tgt:
      return 0
    else:
      return 0.2
  elif dep.tgt == hier.tgt:
    return 0.5
  else:
    return 1
```

An interesting effect of using this technique is that matching elements that do not rely on a marker to indicate untranslated words are generated. These can be of very high quality and catch translation errors such as mistranslated pronouns, verbs in the wrong tense, translations of named entities, etc.

Some additional control of the generation of matching elements is required. The RBMT engine allows for alignment to spurious words on both source and target side and the alignment algorithm in the hierarchical tree extraction module inserts spurious words where needed. Matching elements containing spurious words are filtered out. If matching elements contain nodes where the phrases are only made up of punctuation marks, parentheses and other signs, they are also discarded.

Basing substitution on matching elements from an edit script can result in some untranslated words not being handled. If the system finds any untranslated words in the hierachical tree after substitution, a naive approach where substitution is handled based on an initial alignment has been implemented. Lexical substitution was chosen, because much of the structural information is lost during the first substitution phase and using the leaves technique substituted too many nodes, including nodes inserted during the first substitution phase.

These matching elements will be noisy. They still rely on the segmentation of source phrases and noisy source to target language alignment. The RBMT engine can also produce an inaccurate translation, which makes the substitution counter-productive. The problem is to decide which matching elements to use and which to discard. Further refinement of the criteria for generating matching elements can restrict the generation to matching elements of high quality only. This can be too restrictive and result in some of the most interesting matching elements not being generated or no refinement at all. To be able to make use of these interesting matching elements, but avoid some of the noise, all permutations of applying substitutions based on the generated matching elements are created. The new hypotheses are subsequently re-ranked and the highest scoring hypothesis chosen as the translation.

4.5 Re-ranking

A re-ranking module has to find the best hypothesis from an n-best list of hypotheses. Re-ranking has the advantage that it is possible to use complex language models that analyse a larger context than the previous n tokens. Including complex LMs into the decoding process is not always possible as the decoding process is incremental. LMs are also queried many times during decoding and using complex LMs would increase decoding time. It is desirable to use a more powerful model that can model e.g. long-distance dependencies, because the unconstrained application of substitution, and to a lesser degree the controlled version, can break these dependencies and standard ngram models might not penalise this.

4.5.1 Phrase-based re-ranking

In SMT, re-ranking is usually based on ngrams. A ngram language model or combination of language models are used to rescore translation hypotheses. It is possible to combine a standard ngram model with e.g. a cache model, trigger model, conditional random fields model, etc.

A cache model [Kuhn and De Mori, 1990] increases the probability of the current token if a token with the same surface form occurred in the history of the current token. This is a simple model that can model syntactic relations in constructions such as **as** *good* **as**. The probability mass contributed by the cache model is given in (4.1).

$$P(w_i|h) = \frac{1}{M} \sum_{j=1}^{M} \delta(w_i|w_{i-j}) \tag{4.1}$$

M is the length of the history, w_i is the current word, w_{i-j} is a word in the history and $\delta(w_i|w_{i-j})$ increases by one every time a token with the same surface form occurs in the history. The basic idea is that if a word has ocurred in the history, it is probable that it will occur again.

Trigger models [Lau et al., 1993] are a generalisation of the cache model. In the trigger model, the occurrence of a different token can also increase the probability of the current token. Triggers model relationships between a word and words occurring in the history. Such a set is called a *trigger pair*. The length of the history varies between implementations and the modelled relationships can be syntactic, semantic or anything else. Usually, the history will go back to either the preceeding period, newline or paragraph marker. The cache model does not need any training because the correlation is measured for only one token. Trigger models measure correlations to other words and need to be trained. The correlation measure for triggers are calculated as in Equation (4.1) where $\delta(w_i|w_{i-j})$ is a correlation measure such as mutual information.

The use of triggers is appealing from the point of view that it is possible to model relationships that span more than e.g. a 5-gram. For languages with freer word order or discontinuous phrases and the post-editing approach proposed in this thesis, these relationships are interesting for re-ranking. The substitution phase can break these relationships but the score created by a standard language model may not penalise this because long distance dependencies are not modelled. Trigger pairs are also interesting because they present a method to incorporate relationships to the source sentence. In [Lavecchia et al., 2007], trigger models were used to produce a bilingual dictionary to replace the one produced by GIZA++ in a standard SMT pipeline. On the sentence level, the source sentence is concatenated to the target sentence and partial mutual information is computed. At the corpus level, global mutual information is computed and the best triggers according to some threshold are kept.

4.5.2 Re-ranking on hierachical trees

With regards to hierarchical decoding, which produces trees, forest re-ranking is also possible. [Li and Khudanpur, 2009] create hypergraphs. Hypergraphs encode multiple derivation trees that share nodes and vertices and can compactly represent an exponential number of derivation trees. The vertices represent the instantiation of a SCFG rule. Given a hypergraph of derivation trees, the n-best derivations are extracted based on a probability distribution over vertices. This approach does not perform as well as the ngram-based re-ranking approach used for comparison.

4.5.3 Re-ranking approach

The re-ranking approach implemented in the post-editing approach is phrase-based. The structure of the trees after substitution are neither dependency trees nor hierarchical trees and can only be classified as labelled, ordered trees. This makes it difficult to take point of departure in structure. As a consequence, the re-ranking is based on ngrams and the re-ranking is performed by the language model used in SMT decoding. This approach has been used for re-ranking in a multi-engine MT setup where the output of component MT systems were combined into one n-best list [Hildebrand and Vogel, 2008]. Similar to this approach, the re-ranking is independent of the translation system score. After substitution, this score is no longer comparable. This approach does not take advantage of the possiblity to use advanced language models and future work should improve the re-ranking approach.

To prevent the language model from choosing a sentence with OOV words, the baseline translation is not included in the re-ranking.

Chapter 5

Experiments

The experiments will be conducted between Danish and English. Danish is a low-resource language for which the traditional SMT approach, which requires a lot of training data, does not work well. The closeness of the language pair also makes it easier to model the source to target language word alignment.

5.1 Evaluation metrics

The results of the experiments will be reported using BLEU [Papineni et al., 2002], TER [Snover et al., 2006], and METEOR [Banerjee and Lavie, 2005]. The evaluation script[1] computes all the mentioned metrics.

Translation edit rate (TER) is a metric similar to TED. TER is the number of edit operations needed to turn a translation hypothesis into the reference translation normalised over the average lengths of the references. For evaluation of the hybrid system, only one reference is used and TER is normalised over that sentence length.

Bilingual Evaluation Understudy (BLEU) is the de facto standard for reporting the performance of translation systems. The BLEU metric compares the ngram overlap between one or more reference translations and is called modified because each ngram may only match once and instead of computing recall, a brevity penalty is used to modify the metric. Otherwise, short translations with high ngram overlap, but low recall would achieve a high BLEU score. Unfortunately, BLEU is biased towards SMT systems. This is not due to any flaw in the metric, but rather that the minimum error rate training used to learn parameters is set to maximise the BLEU score on a development set and this leads in general to higher BLEU scores.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is designed to overcome some of the problems of BLEU. METEOR

[1] `http://kheafield.com/code/scoring.tar.gz`

produces scores at the sentence level that seem to correlate better with human judgements than BLEU and uses stemming and synonymy matching in addition to exact unigram matching.

The Danish and English test sets from WMT 2008 [Callison-Burch et al., 2008] were used for computing BLEU, TER and METEOR. The output from the hybrid system is tokenised for the evaluation script and capitalisation ignored when computing the evaluation metrics.

5.2 Experimental setup

Two sets of five experiments have been conducted. The first set of experiments used a translation and language model trained on the first 100,000 lines of Europarl. The models in the second experiment was trained using the entire Europarl corpus.

5.2.1 TED skeleton selection

The impact of choosing the translation hypothesis with a minimal edit distance to the dependency tree from the rule-based system is investigated. 3 different settings have been used for the experiment. In one setting, the cost functions adhere to the constrictions of computing a distance metric.

A variant of TED is known as *tree alignment* [Bille, 2005]. Tree alignment corresponds to performing all insertions before all deletions. 2 settings test the impact of biasing the insertion and deletion cost functions to assign a lower cost to inserting/deleting nonterminals i.e. turning the dependency tree into the hierarchical tree and vice versa.

TED is computed for an n-best list of 20 translation hypotheses and the best performing setting reported on.

5.2.2 Graph techniques

An experiment using the leaves technique has been conducted. The necessary information to use relations or children is not present in the structures. To be able to compare a more naive approach, a technique using lexical substitution, i.e. subtree substitution where the subtree is only one node, will also be reported on.

The dependency subtree is transformed before being inserted into the hierarchical tree. To ensure that the surface string created by the newly created tree will have the correct word ordering, an order must be enforced on the nodes. Seen in isolation, the subtree is continuous and the sentence positions, while not comparable to the tree they are inserted in, indicate the relative word order. To create the insertion tree, a dummy node is created and all the dependency nodes are inserted as leaf nodes of this node. The dummy node is inserted before the root node of the hierarchical subtree and is renamed to have the same ids, parent, etc. Subsequently, the hierarchical nodes are removed from the tree.

The experiment is performed using the best hypothesis and using TED to chose the skeleton. The best performing setting will be reported on.

5.2.3 TED and re-ranking

An experiment where the mappings that represent a rename operation, which are produced during TED computation, are extracted and used as matching elements is conducted. All mapping elements containing the root node of either tree are discarded as well as matching elements consisting only of parentheses and punctutation. All combinations of substitutions based on the extracted matching elements are performed, except those that are not possible because a node in a matching element have already been replaced by a larger subtree.

The extracted matching elements may not incorporate all the untranslated nodes. To take this into account, all untranslated nodes are subsequently translated using lexical substitution as mentioned above. The subtrees inserted into the hierarchical tree will undergo the same transformation as the subtrees inserted using the leaves technique.

Multiple derivations will result in the same surface form, but different structures. This experiment is evaluated using both the 1-best hypothesis as skeleton and choosing the skeleton using TED to study the impact on performance by choosing the most similar structure rather than the best translation. All three settings from Section 5.2.1 are tested and the best performing experiment reported on.

5.3 Automatic evaluation

The results of the automatic evaluation can be seen in Tables 5.1 and 5.2. *TED-R* is the unconstrained substitution approach based on matching elements extracted from edit scripts and reranking of hypotheses. *Skeleton* indicates that TED was used to pick the hierarchical tree. The best evaluations are in bold.

5.3.1 100k experiments

The RBMT baseline is outperformed by all hybrid configurations, though it does have a higher METEOR score than the SMT baseline and experiment 3. Experiment 1 obtains the best BLEU score with an increase of 2.65 BLEU points. Experiment 5 obtains an increase of 2.55 BLEU points and improves TER and METEOR with 2.66 and 4.15 points respectively over the best baseline scores.

Metrics:	BLEU	TER	METEOR
RBMT baseline	19.35	64.54	53.19
SMT baseline	22.63	63.10	50.72
Lexical substitution	**25.28**	60.56	57.24
Leaves technique	21.96	64.80	54.32
TED skeleton(any bias)	22.63	62.98	50.75
TED-R 1-best	25.16	60.51	57.31
TED-R skeleton(any bias)	25.18	**60.44**	**57.34**

Table 5.1: Automatic evaluation of hybrid system trained on 100k lines of Europarl.

Metrics:	BLEU	TER	METEOR
RBMT baseline	19.35	64.54	53.19
SMT baseline	30.16	57.16	59.51
Lexical substitution	**30.53**	**56.40**	**61.22**
Leaves technique	29.06	57.96	60.09
TED skeleton(any bias)	30.16	57.08	59.46
TED-R 1-best	29.78	57.25	59.87
TED-R skeleton(any bias)	29.99	56.72	60.79

Table 5.2: Automatic evaluation of hybrid system trained on the Europarl corpus.

Experiment 2 outperforms the RBMT system in terms of BLEU and METEOR, but does not improve over the SMT baseline in BLEU and TER metrics.

5.3.2 Europarl

In all automatic evaluation metrics, the RBMT module underperforms the SMT baseline. All experiments with the hybrid system outperform the RBMT baseline and in terms of METEOR the experiments also outperform the SMT baseline, with the exception of experiment 3.

The scores achieved when using metric TED to select the hierachical tree, and with no substitution phase, resulted in a decrease in BLEU compared to the SMT baseline by 0.15. Biasing the cost functions resulted in the same BLEU score as the SMT component. The score is the same irrespective of how the insertion and deletion operations were biased, suggesting that to a great extent, the same matching elements were produced. A small difference in TER and METEOR indicates that the alignment was not identical.

Experiment 1 outperforms all other experiments by 0.4 BLEU, 0.3 TER and 0.5 METEOR to the next best scoring experiments. None of the translated words in the existing translation hypothesis are changed — only OOV

Table 5.3: Three examples from the test data.

SMT baseline	Udfordringen from reunification is perhaps even...
Lexical substitution	The challenge from reunification is perhaps even...
Leaves	The challenge from the reunification is perhaps even...

Table 5.4: TED-R output, SMT baseline and reference translation.

Reference	The fundamental notion of the general interest ought, moreover, to spur us all to approve the proposal to include recitals relating to employment in the regulations and decisions on the mergers currently dictated by considerations of competitiveness.
SMT baseline	in fact Elementær general economic interest should encourage us all to approve the proposal to include employment considerations in the regulations and in **fusionsbeslutningerne today sets out of competition .**
TED-R	in fact Elementary general economic interest should encourage us all to approve the proposal to include labour market considerations in the regulations and in **the merger decisions currently set by competition .**

words are replaced. It follows that, supposing that the ngram replacing the OOV word is not longer in length, the evaluation metrics will not decrease. There is only the possibility of increasing the score if the inserted word is present in the reference translation.

This will not improve translation errors made by the SMT system and while the automatic metrics increase for this approach it does not use the strengths of rule-based systems. It is essentially an extension where an additional dictionary is used to handle OOV words.

Still taking point of departure in OOV words, the leaves technique employed in experiment 2 tries to improve the translation by also replacing the surrounding words. When encountering OOV words, the surrounding translation choices are more uncertain. An example of the approach can be seen in Table 5.3.

Udfordringen is correctly substituted by *The challenge* in both experiments. However, lexical substitution does not insert the determiner *the* before *reunification* as it should in this example.

Experiment 5 produces interesting results. An example of the output can be compared with the SMT baseline and the reference in Table 5.4.

Besides substituting *Elementær* which is on OOV word, *employment* is changed to *labour market*. That is an example of the interesting substitutions that are possible when matching elements are extracted from a TED edit script. In this case, the substitution will decrease the BLEU score because the words *labour* and *market* are not in the reference translation while *employment* is. On the other hand, *the, merger, decisions, currently* and *by* are all present in the reference and increases the BLEU score, when inserted into the sentence.

Manual inspection of the sentences selected using re-ranking shows that the most accurate translation is not always selected by the re-ranking module, but experiment 5 still significantly outperforms the experiment 2 in BLEU.

5.4 Manual evaluation

The evaluators are shown 20 sentences randomly extracted from the test set. They are asked to rank the sentences on a scale from 1-5 with 5 being the highest and it is possible to assign the same score to multiple translation alternatives. This evaluation was inspired by the sentence ranking evaluation in [Callison-Burch et al., 2007]. The five sentences will come from the RBMT and SMT baselines, lexical substitution, leaves approach and TED-R skeleton.

Two evaluations were carried out. The first evaluation campaign targeted Danes who have studied translation with English as the second language. The second targeted native English speakers.

5.4.1 Danish translators

The evaluators have completed a bachelors degree in translation with studies in translation of legal, economic and cultural texts. Some of the evaluators also study translation at master level.

For each sentence, the evaluators are shown the Danish source sentence as a reference.

5.4.2 Native English speakers

For English native speakers, the English reference sentence is used as reference.

5.4.3 Results of manual evaluation

5 translators and 3 native speakers participated in the manual evaluation. The outcome of the evaluation campaign can be seen in Table 5.5. The first column contains the systems included in the evaluation and the remaining columns contain the number of times a system received the ranking heading that column. For each rank, the highest number of assignments of that rank is shown in bold, i.e. the SMT baseline received most assignments of rank *1*, so **52** is in bold.

Table 5.5: Rankings from the manual evaluation.

System	1	2	3	4	5
SMT baseline	**52**	**64**	30	12	1
RBMT baseline	14	48	61	29	8
Lexical substitution	3	33	**63**	**58**	3
Leaves	6	33	61	55	5
TED-R	3	35	46	55	**21**

The baseline systems receive many low rankings, especially the SMT baseline. The configurations of the hybrid systems do not receive as many low rankings and the baselines make up 85% of the lowest ranking. The distribution between systems is more even for the second lowest ranking with the baselines only accounting for 52.6%, but still more than the hybrid systems. The distribution changes for the middle ranking. The top scorer is lexical substitution with a small margin to the RBMT baseline and the leaves approach, while the SMT system received fewest. There are by far most rankings of *3* in the evaluation. This could indicate that many of the translations produced are acceptable to use for gisting, i.e. get an impression of what information the source text conveys, but not enough to give a complete understanding. It can also be a result of being the middle value and chosen when the evaluators are otherwise in doubt.

Lexical substitution is also the top scorer in the second-best ranking, followed closely by the other hybrid configurations. The downward trend from for the second-lowest and middle scores is continued by the SMT baseline. The number of rankings for GramTrans also decreases by half which means that the hybrid systems account for 80.3% of the second best rankings.

TED-R recieves more top rankings than the other systems combined, accounting for more than half of the scores (55.3%). The RBMT baseline achieves second most top-rankings. This can be attributed to the cases where the rules were succesful in disambiguation, transfer and generation and created very accurate translations, as is the hallmark of RBMT.

Chapter 6

Discussion

The noisy alignments which form the basis for substitution introduce errors. Among them repetition of names, prepositions and numbers, which are the same errors observed using SPE and in [Federmann et al., 2010]. In some cases, instead of seeing an improvement in the structure of the sentence, the structure degrades. This is especially clear when handling numbers. And of course, there are cases where the RBMT system produces an inaccurate translation which — despite using re-ranking — is inserted into the translation hypothesis.

Despite all these problems, the evaluation indicates that the post-editing approach improves translation accuracy.

6.1 Results

The RBMT baseline underperforms all other MT systems, except for two baselines in the 100k set of experiments where the statistical model is weak. It follows that inserting subtrees from the RBMT baseline should reduce BLEU, TER and METEOR scores in the cases where surface structure in the hierarchical tree is replaced with RBMT surface structure.

The improvements evident in the 100k experiments indicate that the hybrid approaches succesfully finds appropriate RBMT subtrees to insert into the hierarchical skeleton. While the substitution of OOV words improve the metrics, the succes of the implementations of the post-editing approach differs widely. This can indicate that the approach replaces SMT surface structure that is present in the reference translation. When there is little or no degredation in the metrics, it suggests that correct material from the RBMT output is inserted into the hierarchical skeleton.

The fact that lexical substitution achieves a significant increase in all metrics in both sets of experiments is not surprising. The approach finds OOV words and translates them using the RBMT lexicon while leaving the surrounding surface forms untouched. Because of the noisy matching ele-

SMT baseline	Hertil at the same time , however , we must add that we must draw omsorg for the administrative beredskab .
Lexical substitution	Here at the same time , however , we must add that we must draw care for the administrative readiness .
Leaves	Here , however , we must add that we at the same time must care draw for the administrative readiness .

Table 6.1: Degradation of surface structure. (100k)

SMT baseline	I Lad us hope that in the next few hours can bibringe you some better nyheder . the can I unfortunately not at the moment .
Lexical sub.	I Let us hope that in the next few hours can in you some better the . the can I unfortunately not at the moment .
Leaves	Let us hope , that I can give you some better news in the coming hours . the can I unfortunately not at the moment .

Table 6.2: Improvement of surface structure. (100k)

ments, it will sometimes improve the translation and sometimes noise or wrong words will be introduced, but the introduction of noise and wrong words is not detrimental for the metrics as the penalty incurred for untranslated words and wrongly translated words is the same — assuming that the number of tokens is similar. Subtree substitution also relies on an alignment that is computed based on the initial alignment. Lexical substitution does not need further refinement and can avoid the potential sources of errors that occur at later processing steps.

When comparing the evaluation of the leaves technique and lexical substitution, the automatic evaluation of both sets of experiments indicates a substantial gap in translation accuracy and subtree substitution consistently underperforms lexical substitution. Inspection of the translations show that substitution based on the leaves technique can degrade the structure even though the correct words are still present in the translation as can be seen in Table 6.1.

The leaves technique also improves the translations produced by the SMT baseline. Table 6.2 shows an example where the inserted subtree both correctly translates the OOV word and rearranges the surface structure to be more fluent.

Improvements such as the one in Table 6.2 will not be visible in the automatic metrics because of the degradation that occurs in cases such as in Table 6.1 and when the inserted surface structure is not present in the reference translation.

Manual evaluation of a sample from the experiments in Section 5.3.2 shows a high correlation between the two methods. The fact that lexical substitution and subtree substitution perform almost identically in the manual evaluation, but there is a difference of 1.5 BLEU and TER points and 1 METEOR point, suggests that the inserted material usually does not appear in the reference translation. It might also indicate that the isolation of subtrees for substitution works well, but the deterioration/improvement of the structure is minimal compared to lexical substitution when the statistical models are more powerful.

The manual evaluation reveals a distribution where the majority of the rankings for the baselines are in the lower half. The distribution of the rankings for the hybrid systems tend more towards the mid-to-upper rankings, with TED-R having more of the distribution around the second-best and highest score. If the sample questions used for manual evaluation accurately represent the test data, the results indicate that the hybridisation approach succesfully uses the strengths of the component systems to create more accurate translations. A substantial decrease in assignment of the lowest ranking is observed as well as a shift in the distribution towards average and above average rankings.

Consistent with the automatic evaluation, TED-R is ranked higher than subtree substitution using the leaves technique. However, manual evaluation also ranks TED-R higher than lexical substitution. The reason for this discrepancy is found in the inserted words from the RBMT system which are not present in the reference, but which the evaluators assign a high rank when inserted correctly into the skeleton. While lexical substitution can find a word that appear in the reference translation and insert it in the hybrid translation hypothesis, the method does not change the surrounding surface forms. In some cases, this results in disfluent surface structures which the evaluators penalise while TED-R changes the surrounding surface structure as well, which can create a syntactically well-formed structure that the evaluators reward in the evaluation.

As mentioned in Section 5.2.3, the translation alternatives created by TED-R are post-processed. The computation and extraction of matching elements from the edit distance mapping might not entail matching elements for all OOV words. Lexical substitution is applied before the re-ranking step which means that there will be a translation hypothesis created using only lexical substitution in the set of translation hypotheses that are being re-ranked. The automatic metric scores for lexical substitution and TED-R are quite similar. To verify that the re-ranker does not primarily choose the translation hypothesis created using lexical substitution, an experiment with the lexical substitution post-processing step removed was conducted. The experiment achieved scores of 29.97 BLEU, 56.76 TER, 60.78 METEOR and 25.15 BLEU, 60.44 TER, 57.25 METEOR on the models trained on the full Europarl corpus and 100,000 lines of Europarl respectively. The discrepancy to the scores in Table 5.2 are not significant and indicates that the rename cost function defined in Listing 4.2 creates matching elements that incorporate almost all OOV words.

SMT baseline	(COM (1999) 493 - C5-0320 / 1999 - 1999 / 2208 (COS))
Leaves	(came (1999) 493 - C5-0320/1999-1999/2208 (COM COS)) - C5-0320 / 1999 - 1999 / 2208 (
TED-R	(COM (1999) 493 - C5-0320/1999-1999/2208 / 1999 - 1999 / 2208 (COS))

Table 6.3: Substitution of numbers.

6.2 Problems

6.2.1 OOV words

Generally, the evaluators penalised the occurences of untranslated words. These occurences seemed to cluster around the initial word in the translated sentence. An inspection of the derivations in the trace information has not been conducted and the reason is currently unknown. It is however certain that these words are in the phrase table.

6.2.2 Subtree Matching

Subtree substitution, whether using leaves or TED-R, does not handle parentheses, hyphens and numbers well. The structure severely degrades when performing substitution near these environments. The example in Table 6.3 shows the errors made by the substitution algorithm. An entire subphrase is duplicated using the leaves technique which introduces an opening parenthesis with no closing counterpart and includes the erroneous translation *came*, while TED-R duplicates */ 1999 - 1999 / 2208*.

The reason for these wayward substitutions can be found in the dependency tree. A simplified version of the dependency tree that contains the phrase from Table 6.3 can be seen in Figure 6.1. The matching parentheses are not part of the same subtree and this is the root cause of the problem. The leaves technique is very sensitive to these errors and there is no easy way to prevent spurious parentheses from being introduced. Re-ranking in TED-R could filter these hypotheses out, but because the re-ranking module cannot model this dependency, the sentences with these errors are not always discarded.

In the manual evaluation campaign, the sentence from Figure 6.1 was included in the sample sentences. It would seem that the many evaluators did not view this error as important or it was ignored. Translation including numbers is important and cannot be disregarded. It would be impossible to find the Council decision based on the translations and dates or monetary amounts might change drastically, which would not be acceptable if the translated text should be ready for publishing after translation. For gisting,

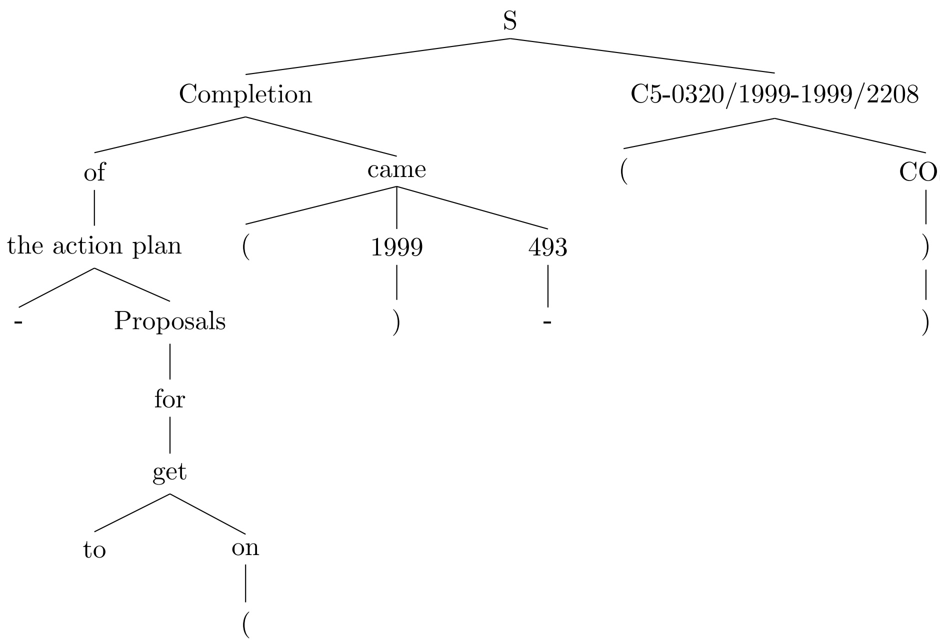

Figure 6.1: Simplified RBMT tree.

where the user knows that the translation is not perfect, this may constitute less of a problem.

As mentioned in section 4.4, the system is sensitive to the alignment quality. The initial alignment is based on the source to target language alignment. In the RBMT module, it is mostly word-based except for simplex phrases and in Moses the alignment must be recomputed due to the simplicity of the modification to the Moses decoder which cannot output the word alignment. The modelling only handles alignment crossing one non-terminal. The alignment algorithm reduces all alignment problems to these cases by assuming a weak reordering model. This creates wrong alignments and introduces a lot of noise. To improve the subtree substitution and overall performance, the initial alignment is an obvious place to start as the generation of matching elements rely on this alignment. Even TED relies on this alignment because the source and target phrases are treated as labels upon which the cost functions for edit operations are defined.

6.2.3 Insertion problems

The implementation encounters some problems when inserting subtrees where the identified root node of the subtree is also the root node of the entire hierarchical tree. The arrangement of the nodes in the subtree are still handled correctly, but the insertion can go astray and insert the insertion tree on the wrong side of a sibling node, which degrades the structure or is sometimes construed as syntactic transformations where an adverbial phrase is e.g. topicalised.

6.3 Future work

6.3.1 Languages and formalisms

The chosen languages are closely related Germanic languages. While the results seem promising, the applicability of the approach should be tested on a more distant language pair, e.g. Chinese-English or Russian-English if you wish to preserve the possibility of using METEOR for evaluation, but any distant pair for which an RBMT system exists can be used — provided a tree output is available.

The implementation substitutes dependency subtrees into a hierarchical CFG-style tree. A second test of the hybridisation approach is to combine systems where the structures are not as diverse. Hierarchical systems are derived from a SCFG so a RBMT system that outputs CFG trees such as LUCY, could be used to test the generality of the hybridisation approach. As the TED-R approach does not rely on markers for OOV words, an implementation where hierarchical subtrees are inserted into the RBMT output should also be conducted. The problem of inserting CFG-style subtrees into a dependency tree and generating the correct surface structure must be resolved or a different RBMT system which produce CFG-style trees implemented.

The implementation of the leaves technique relies on the diversity of the tree structures, i.e. that there are element-level similarities between hierarchical leaf nodes and both terminal and nonterminal dependency nodes and that the subtree rooted in a dependency node can be aligned to a hierarchical subtree. The refinement method would have to be altered. The other approaches — relations and children — would be good candidates if both tree structures were dependency trees or linguistic syntax trees with phrasal categories.

A change of formalism would not require alterations of the tree edit distance approach, as long as the structures are in fact tree structures.

6.3.2 Re-ranking

The re-ranking module is based on a classical n-gram language model. As discussed earlier, the ngram language model cannot model long-distance dependencies beyond e.g. 5 tokens, as it is a 5-gram language model. The re-ranking could be made more fine-grained by interpolating e.g. triggers. Training a larger language model using more monolingual data than is available in Europarl would also result in a more powerful re-ranking module or, since the methods are not incremental, a factored language model could be used.

Chapter 7

Conclusion

The post-editing approach proposed in this thesis combines the strengths of statistical and rule-based machine translation and improve translation accuracy, especially for the least accurate translations.

Automatic evaluation shows an increase in BLEU, METEOR and TER over the baselines for lexical substitution and for the approach based on tree edit distance and re-ranking.

Manual evaluation on a sample of the test data shows that the hybrid translations were generally ranked higher, indicating that the hybrid approach produces more accurate translations.

Reuse of the word alignment in Moses and an improved re-ranking module could further improve the results gained from the post-editing approach. Alignment to other structures with this approach needs to be investigated as well as the application of the post-editing approach to more distant language pairs. The current implmentation is knowledge-poor and can in principle be applied to any langauge pair, but it is also possible to extend the approach and enrich the structures with language-dependent information, e.g. parts of speech, lemma, etc.

The reported results indicate that HMT can be used effectively to improve MT for low-resource languages. The approach does not solve the main problems of MT, we still need better models of translation that are based on linguistic insights, but the approach can help lower the language barrier, facilitate information access and enhance the free flow of information.

Bibliography

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 65.

[Bick, 2000] Bick, E. (2000). The parsing system PALAVRAS. *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Arhus.*

[Bick, 2001] Bick, E. (2001). En constraint grammar parser for dansk. *Peter Widell & Mette Kunøe (eds.)*, 8:40–50.

[Bick, 2007] Bick, E. (2007). Dan2eng: Wide-coverage danish-english machine translation. *Proceedings of Machine Translation Summit XI*, pages 37–43.

[Bick, 2009] Bick, E. (2009). Introducing probabilistic information in constraint grammar parsing. In *Proceedings of Corpus Linguistics 2009, Liverpool, UK.* Citeseer.

[Bille, 2005] Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical computer science*, 337(1-3):217–239.

[Brants and Franz, 2006] Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1.

[Brown et al., 1990] Brown, P., Cocke, J., Pietra, S., Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

[Brown et al., 1993] Brown, P., Pietra, S., Pietra, V., and Mercer, R. (1993). The mathematic of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

[Callison-Burch et al., 2007] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics.

[Callison-Burch et al., 2008] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. Citeseer.

[Chen and Eisele, 2010] Chen, Y. and Eisele, A. (2010). Integrating a rule-based with a hierarchical translation system. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*.

[Chiang, 2005] Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

[Eisele et al., 2008] Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., and Chen, Y. (2008). Hybrid machine translation architectures within and beyond the EuroMatrix project. In *Proceedings of the 12th annual conference of the European Association for Machine Translation (EAMT 2008)*, pages 27–34.

[Eurobarometer, 2011] Eurobarometer (2011). User language preferences online, (flash no. 313). Electronically published at `http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf`. Directorate-General for Communication, European Comission.

[Federmann et al., 2010] Federmann, C., Eisele, A., Uszkoreit, H., Chen, Y., Hunsicker, S., and Xu, J. (2010). Further experiments with shallow hybrid mt systems. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 77–81. Association for Computational Linguistics.

[Hildebrand and Vogel, 2008] Hildebrand, A. and Vogel, S. (2008). Combination of machine translation systems via hypothesis selection from combined n-best lists. In *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 254–261. Citeseer.

[Hutchins, 2007] Hutchins, J. (2007). Machine translation: A concise history. *Computer aided translation: theory and practice. Chinese University of Hong Kong.*

[Hutchins and Somers, 1992] Hutchins, W. and Somers, H. (1992). *An introduction to machine translation.* Academic Press New York.

[Karlsson, 1995] Karlsson, F. (1995). *Constraint Grammar: a language-independent system for parsing unrestricted text.* Walter de Gruyter.

[Koehn, 2005] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5. Citeseer.

[Koehn, 2010] Koehn, P. (2010). An experimental management system. *The Prague Bulletin of Mathematical Linguistics*, 94(-1):87–96.

[Koehn and Hoang, 2007] Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, volume 868, page 876.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

[Koehn and Monz, 2006] Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121. Association for Computational Linguistics.

[Koehn et al., 2003] Koehn, P., Och, F., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

[Kuhn and De Mori, 1990] Kuhn, R. and De Mori, R. (1990). A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 570–583.

[Lau et al., 1993] Lau, R., Rosenfeld, R., and Roukos, S. (1993). Trigger-based language models: A maximum entropy approach. In *ICASSP*, pages 45–48. IEEE.

[Lavecchia et al., 2007] Lavecchia, C., Smaili, K., Langlois, D., and Haton, J. (2007). Using inter-lingual triggers for machine translation. In *Proceedings of the eighth conference in the annual series of INTERSPEECH, Antwerp, Belgium, August.*

[Li and Khudanpur, 2009] Li, Z. and Khudanpur, S. (2009). Forest reranking for machine translation with the perceptron algorithm. *GALE book chapter on MT From Text.*

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Shvaiko and Euzenat, 2005] Shvaiko, P. and Euzenat, J. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics IV*, pages 146–171.

[Snover et al., 2006] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231. Citeseer.

[Stolcke, 2002] Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904. Citeseer.

[Thurmair, 2009] Thurmair, G. (2009). Comparing different architectures of Hybrid Machine Translation systems. In *Proceedings of the MT Summit XII*, pages 340–347.

[Weaver, 1955] Weaver, W. (1955). Translation. *Machine translation of languages*, 14:15–23.

[Zhang and Shasha, 1989] Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18(6):1245–1262.

Appendix A

Manual evaluation summary

5.5

Sentence 1

Reference: One problem specific to the Czech Republic, especially on the German-Czech border, is the problem of prostitution, especially child prostitution.

Source: Et særligt problem i Tjekkiet, nærmere bestemt ved den tjekkisk-tyske grænse, er prostitution, især børneprostitution.

Et particular problem in the Czech Republic, specifically at the border tjekkisk-tyske is prostitution, particularly child prostitution.

A special problem in the Czech Republic, specifically at the Czech-German border, prostitution, especially child prostitution is.

A particular problem in the Czech Republic, specifically at the border Czech-German is prostitution, particularly child prostitution.

A particular problem in the Czech Republic, specifically at the border Czech-German is prostitution, particularly child prostitution.

A special problem in the Czech Republic and, more specifically, on the border between Czech-German, prostitution, particularly child prostitution.

Sentence 2

Reference: And, in addition, we really must see to it at the same time that the proper administrative practices are in place.

Source: Hertil må man dog tilføje, at man samtidig må drage omsorg for det administrative beredskab.

Hertil we must add, however, that we must, at the same time, ensure the administrative preparedness.

Here you must however add that you must take care for the administrative readiness at the same time.

Here we must add, however, that we must, at the same time, ensure the administrative preparedness.

Here we must add, however, that we must, at the same time, ensure the administrative preparedness.

Here we should, however, like to add that, at the same time, it must ensure the administrative preparedness.

Sentence 3

Reference: Mr President, while the European food chain is one of the safest in the world, poor practice, and indeed recent scandals, have set an agenda to which our Commissioner, Mr Byrne, has responded admirably.

Source: Hr. formand, skønt den europæiske fødekæde er en af de sikreste i verden, har dårlig praksis samt de seneste skandaler sat en dagsorden, som kommissær Byrne har håndteret på beundringsværdig vis.

Mr President, although the European food chain is one of the safest in the world, bad practices, as well as the recent scandals have set an agenda, as Commissioner Byrne has handled admirably.

Mr chairman, although the European food chain is one of the safer ones in the world, bad practice as well as the latest scandals have established an agenda that a commissioner Byrne has handled in an admirable way.

Mr President, although the European food chain is one of the safest in the world, bad practices, as well as the recent scandals have set an agenda, as Commissioner Byrne has handled admirably.

Mr President, although the European food chain is one of the safest in the world, bad practices, as well as the recent scandals have set an agenda, as Commissioner Byrne has handled admirably.

Mr President, although the European food chain is one of the safest in the world, bad practices, as well as the latest scandals have set an agenda that Commissioner Byrne has demonstrated admirably.

Sentence 4	
Reference:	They are not leader groups but reflect a diversity which, with 30 Member States, will only increase.
Source:	Det drejer sig ikke om førerfelter, men om en mangfoldighed, som kun bliver større med 30 medlemsstater.
	It is not a question of førerfelter, but on a diversity which will only be increased by 30 Member States.
	This isn't about leader fields, but about a diversity, which only becomes greater with 30 member states.
	it is not a question of leader fields, but on a diversity which will only be increased by 30 Member States.
	it is not a question of leader fields, but on a diversity which will only be increased by 30 Member States.
	It is not a question of leader fields, but about a diversity which only increase by 30 Member States.
Sentence 5	
Reference:	If I am being honest, I feel we are to some extent deceiving the electorate.
Source:	
	Det ligner ærligt talt vælgerbedrag.
	The resembles frankly vælgerbedrag.
	That honestly resembles spoken electorate cheating.
	The resembles frankly electorate cheating.
	The resembles spoken electorate cheating.
	Quite frankly, the resembles electorate cheating.
Sentence 6	
Reference:	Moreover the development of nuclear energy and carbon sinks must be excluded from the calculations for the reduction of emissions.
Source:	I øvrigt skal udviklingen af atomenergien og kulstofdræn tages ud af udregningerne for emissionssænkningerne.
	In addition, the development of nuclear energy and carbon sinks must be taken out of calculations for emissionssænkningerne.
	Otherwise the development of the nuclear energy and carbon drains is to be taken out of the calculations for the emission reductions.
	in addition, the development of nuclear energy and carbon sinks must be taken out of calculations for the emission reductions.
	in addition, the development of nuclear energy and carbon sinks must be taken out of calculations for the emission reductions.
	You must also the development of nuclear energy and carbon sinks be taken out of calculations for the emission reductions.

Sentence 7

Reference: Moreover, volume and number of retail payments only form a small percentage of the total number of credit transfers.

Source: Endvidere er volumen og antallet af detailbetalinger kun en ubetydelig procentdel af de samlede overførsler.

Endvidere is volume and the number of retail payments only a small percentage of the total transfers.

Furthermore volume and the number of retail-payments are only an insignificant percentage of the overall transfers.

Furthermore is volume and the number of retail payments only a small percentage of the total transfers.

Furthermore is volume and the number of retail payments only a small percentage of the total transfers.

Furthermore volume and the number of retail-payments is only an insignificant percentage of the total transfers.

Sentence 8

Reference: Regardless of the reasons that may have caused this mistake, it is certain that it is motivated by the constant initiatives, originating in the Commission, that are carried out with regard to legislation.

Source: Uanset årsagerne til denne fejl er det helt sikkert, at den er begrundet i den strøm af initiativer, som Kommissionen tager med hensyn til lovgivning.

Uanset the causes of this mistake, there is no doubt that it is justified by the flow of initiatives that the Commission take with regard to legislation.

Regardless of the causes of this mistake it's totally safe that it's reasonable in that stream of initiatives that the Commission takes concerning legislation.

Regardless of the causes of this mistake, there is no doubt that it is justified by the flow of initiatives that the Commission take with regard to legislation.

Regardless of the causes of this mistake, there is no doubt that it is justified by the flow of initiatives that the Commission take with regard legislation.

Regardless of the causes of this mistake, there is no doubt that it is based on it flow of initiatives that the Commission take with regard to legislation.

Sentence 9

Reference:	(A5-0313/2000) by Mr Marinho, on behalf of the Committee on Citizens' Freedoms and Rights, Justice and Home Affairs, on the initiative of the French Republic with a view to adopting a Council Framework Decision on money laundering, the identification, tracing, freezing, seizing and confiscation of instrumentalities and the proceeds from crime [10232/2000 - C5-0393/2000 - 2000/0814(CNS)],
Source:	A5-0313/2000 af Marinho for Udvalget om Borgernes Friheder og Rettigheder og Retlige og Indre Anliggender om initiativ fra Den Franske Republik med henblik på vedtagelse af Rådets afgørelse om hvidvaskning af penge, identifikation, opsporing, indefrysning eller beslaglæggelse og konfiskation af redskaber og udbytte fra strafbart forhold (10232/2000 - C5-0393/2000 - 2000/0814(CNS));
	A5-0313 / 2000) by Mr Marinho, on behalf of the Committee on Citizens' Freedoms and Rights, Justice and Home Affairs, on the initiative of the French Republic with a view to the adoption of a Council decision on money laundering, identification, tracing, freezing or seizing and confiscation of tools and the proceeds from crime (10232 / 2000 - C5-0393 / 2000 - 2000 / 0814 (CNS)) ;
	A5-0313/2000 of Marinho for Udvalget om Borgerne's Freedoms and Rights and Legal and Inner Matters about initiative from The French Republic in preparation of the passing of the Council's decision about the laundering of money, identification, tracking down, freezing or seizure and confiscation of tools and profit from punishable condition (10232/2000 - C5-0393/2000-2000/0814 (CNS));
	A5-0313/2000 / 2000) by Mr Marinho, on behalf of the Committee on Citizens' Freedoms and Rights, Justice and Home Affairs, on the initiative of the French Republic with a view to the adoption of a Council decision on money laundering, identification, tracing, freezing or seizing and confiscation of tools and the proceeds from crime (10232/2000 / 2000 - C5-0393/2000-2000/0814 / 2000 - 2000 / 0814 (CNS)) ;
	Freedoms and and Rights and and Legal and Inner Inner Matters about about initiative from The French Republic in preparation of the passing of the Council's decision French about about Republic the laundering of money, identification identification, tracking down, freezing freezing or seizure seizure and confiscation confiscation of tools and, profit profit from punishable condition ((10232/2000 - C5-0393/2000-2000/0814 (CNS CNS)) ; 10232 / 2000 C5-0393 / / 2000 2000 - 0814
	A5-0313/2000 / 2000) by Mr Marinho, on behalf of the Committee on Citizens' Freedoms and Rights, Justice and Home Affairs, on the initiative of the French Republic with a view to the adoption of a Council decision on money laundering, identification, tracing, freezing or seizure and confiscation of tools and the proceeds from crime (10232/2000 / 2000 - C5-0393/2000-2000/0814 / 2000 - 2000 / 0814 (CNS)) ;

Sentence 10

Reference: Without a certain degree of cooperation, it is simply not possible for us to uphold a law-governed society in an EU with freedom of movement.

Source: Uden et vist samarbejde kan vi ikke opretholde retssamfundet i et EU med fri bevægelighed.

Uden some cooperation we will not be able to maintain the rule of law in an EU of free movement.

Without a certain cooperation we cannot maintain the community founded on the rule of law in an EU with freedom of movement.

Without some cooperation we will not be able to maintain the rule of law in an EU of free movement.

Without a certain certain cooperation we will not be able to maintain the rule of law in an EU of free movement.

Without a certain cooperation we will not be able to maintain the rule of law in a European Union with freedom of movement.

Sentence 11

Reference: I am sure the Convention members that are to speak here will be able to confirm that.

Source: Jeg er sikker på, at dette kan bekræftes af de medlemmer af Europa-Parlamentet, som har deltaget i forsamlingen.

I am sure that this is confirmed by the members of Europa-Parlamentet who have taken part in this House.

I am certain that this can be confirmed by those members of the European Parliament who has participated in the gathering.

I am sure that this is confirmed by the members of the European Parliament who have taken part in this House.

I am sure that this is confirmed by the members of the European Parliament , who has participated in the gathering

I am sure that this is confirmed by the members of the European Parliament who participated in the House.

Sentence 12

Reference: I should like on behalf of our group to offer express thanks not only to all the members of our group but, most of all, to Vice-President Ingo Friedrich, who coordinated our work.

Source: På vores gruppes vegne vil jeg udtrykkeligt gerne takke ikke bare alle kollegerne i vores egen gruppe, men især Ingo Friedrich, Parlamentets næstformand, fordi han koordinerede vores arbejde.

On behalf of our group, I would expressly like to thank all my fellow Members not only in our own group, but, in particular, Mr Ingo Friedrich, Vice-President of Parliament, because he coordinated our work.

On our group's behalf I'd like to explicitly thank all the colleagues not just in our own group, but especially Ingo Friedrich, the Parliament's vice-chairman, because he coordinated our work.

On behalf of our group, I would expressly like to thank all my fellow Members not only in our own group, but, in particular, Mr Ingo Friedrich, Vice-President of Parliament, because he coordinated our work.

On behalf of our group, I would expressly like to thank all my fellow Members not only in our own group, but, in particular, Mr Ingo Friedrich, Vice-President of Parliament, because he coordinated our work.

On behalf of our group, I would expressly like to thank all my fellow Members not only in our own group, but especially Mr Ingo Friedrich, Vice-President of Parliament, because he coordinated our work.

Sentence 13

Reference: This Charter on social rights is a step backwards compared to the laws of many of the Member States and, thanks precisely to the referrals to these laws, it will serve as a tool to alter them or limit their effectiveness.

Source: Dette charter om sociale rettigheder er en forringelse i forhold til mange medlemsstaters lovgivninger, og netop fordi der hele tiden sker udsættelser, vil chartret være et redskab til at ændre lovgivningerne eller gøre dem ringere.

Dette Charter of social rights are a step backwards compared with many Member States' legislation, and precisely because there are delays, the Charter will be a tool for all the time to change legislation or make them worse.

This charter about social rights is a reduction concerning many member states' legislations, and precicely because all the time there happen postponements, the charter will be a tool to alter the legislations or make them more inferior.

This Charter of social rights are a step backwards compared with many Member States' legislation, and precisely because there are delays, the Charter will be a tool for all the time to change legislation or make them worse.

This Charter of social rights are a step backwards compared with many Member States' legislation, and precisely because there are delays, the Charter will be a tool for all the time to change legislation or make them worse.

This Charter of social rights are a step backwards compared with many Member States' legislations and, precisely because there are delays, the Charter will be a tool to amend legislation or make them more inferior all the time.

Sentence 14

Reference: Mr President, the proposal and adoption of the draft Charter of Fundamental Rights should meet with very broad consensus.

Source: Hr. formand, der burde være meget bred enighed om forslaget og vedtagelsen af udkastet til charter om grundlæggende rettigheder.

Mr President, we should be very broad agreement on the proposal and the adoption of the draft Charter of Fundamental Rights.

Mr chairman, who should be very wide agreement about the proposal and the passing of the draft of charter about fundamental rights.

Mr President, we should be very broad agreement on the proposal and the adoption of the draft Charter of Fundamental Rights.

Mr President, we should be very broad agreement on the proposal and the adoption of the draft Charter of Fundamental Rights.

Mr President, there ought to be very broad agreement on the proposal and the adoption of the draft Charter of fundamental rights.

Sentence 15

Reference: For a start, they must incorporate an amendment to Article 6 of the EU Treaty.

Source:

Allerførst må de optage en henvisning i EU-traktatens artikel 6.

Allerførst they must include a reference in Article 6 EU-traktatens.

First of all they must take up a reference in the EU treaty's article 6.

First of all they must include a reference in Article 6 the EU treaty's.

First of all they must include a reference in Article 6 the EU treaty's.

First of all they must include a reference in the EU treaty's Article 6.

Sentence 16

Reference: Given that only have two options, my decision is unequivocally to vote in favour of the step forward which this Charter represents, however small that step forward may be in terms of the content and scope of the text.

Source: Stillet over for dette valg er min stemme helt klar. Jeg går ind for det fremskridt, som dette charter er, selv om der er tale om et relativt fremskridt både med hensyn til tekstens indhold og betydning.

Stillet vis-à-vis these elections are quite clear in my vote. I am in favour of the progress that this Charter is, even though we are talking about a relatively progress both in terms of its content and meaning.

Put face to face with this choice my voice is quite clear. I support the progress which this charter is, even though there is a relative advance both concerning the text's contents and meaning.

Put vis-à-vis these elections are quite clear in my vote. I am in favour of the progress that this Charter is, even though we are talking about a relatively progress both in terms of its content and meaning.

Put face to face with this choice are quite clear in my vote. I am in favour of the progress that this Charter is, even though we are talking about a relatively progress both in terms of its content and meaning.

Put vis-à-vis these elections are quite clear in my vote. I am in favour of the progress that this charter is, even though there is a relative progress both in terms of its content and meaning.

Sentence 17

Reference: A great deal of headway has been made in the space of a year and the internal dynamics in a number of countries are at long last enhancing the role of the international community, even though the Kosovo crisis is still fresh in the memory.

Source: På et år er der gjort mange fremskridt, og den interne dynamik i en række lande styrker omsider det internationale samfunds rolle, selv om kosovokrisen stadig er i frisk erindring.

In a year much progress has been made, and the internal dynamics in a number of countries forces at long last, the role of the international community, even though the Kosovo crisis is still fresh in our minds.

On a year there have been done a lot of advances, and the internal dynamics in a number of countries finally strengthens the international community's role, even though the Kosovo crisis still is in cheerful memory.

In a year much progress has been made, and the internal dynamics in a number of countries forces at long last, the role of the international community, even though the Kosovo crisis is still fresh in our minds.

In a year much progress has been made, and the internal dynamics in a number of countries forces at long last, the role of the international community, even though the Kosovo crisis is still fresh in our minds.

In a year is that much progress has been made, and the internal dynamics in a number of countries strengthens the role of the international community, even on the Kosovo crisis still still fresh in our minds.

Sentence 18

Reference: More importantly, we must not forget that in numerous countries farmers fatten their piglets themselves and hence have direct access to the compensation fund via the fattened pigs which they sell.

Source: Især må vi heller ikke glemme, at landmændene i mange lande opfeder deres egne smågrise og dermed har direkte adgang til udligningsfonden gennem de slagtesvin, de sælger.

We must Især not forget that farmers in many countries opfeder their own piglets and thus have direct access to udligningsfonden through the slaughter pigs, they sell.

Particularly we mustn't forget either that the farmers in many countries fatten their own piglets and thus direct access to the equalization fund has through the porkers they sell.

We must particularly not forget that farmers in many countries fatten their own piglets and thus have direct access to the equalization fund through the slaughter pigs, they sell.

Particularly we must not forget that the farmers in many countries fatten their own piglets and thus have direct access to the equalization fund through the slaughter pigs, they sell.

Particularly we must not forget that farmers in many countries fatten their own piglets and thus have direct access to the equalization fund through the slaughter pigs, they sell.

Sentence 19

Reference: For this simple reason we have expressed a different viewpoint, and we can only advise this Chamber, on behalf of a majority of the Committee on Legal Affairs and the Internal Market, to accept Amendments Nos 38 or 41 in particular, as these relate to a weakening of export bans.

Source: Af den simple grund har vi her repræsenteret en anden opfattelse, og vi kan på vegne af Udvalget om Retlige Anliggender og Det Indre Marked kun tilråde plenarforsamlingen at følge især forslag 38 eller 41, som beskæftiger sig med en svækkelse af eksportforbuddet.

Af the simple reason we have represented here on behalf of a different view, and we can the Committee on Legal Affairs and the Internal Market only urge the plenary to follow in particular Amendments 38 and 41, which deals with a weakening of the export ban.

For the simple reason we have here represented another perception, and we can on behalfs of the Committe on Legal Affairs and the Internal Market only advise the plenum to follow especially proposals 38 or 41, that concern themselves with a weakening of the export prohibition.

For the simple reason we have represented here on behalf of a different view, and we can the Committee on Legal Affairs and the Internal Market only urge the plenary to follow in particular Amendments 38 and 41, which deals with a weakening of the export ban.

For the simple reason we have represented here a different view, and we can on behalf of the Committee on Legal Affairs and the Internal Market only urge the plenary to follow in particular Amendments 38 and 41, which deals with a weakening of the export ban.

For the simple reason we have represented here a different view, and we can on behalf of the Committee on Legal Affairs and the Internal Market only urge the plenary to follow that in particular Amendments 38 and 41, deals with a weakening of the export ban.

Sentence 20

Reference: Similarly, we need localisation of content production, which is a means of adapting products according to linguistic and cultural needs and the liking of consumers.

Source: Ligeledes er der brug for indholdsproduktionens lokalisering, der er en tilpasning af produkterne til at modsvare sproglige og kulturelle behov samt forbrugernes ønsker.

Ligeledes we need indholdsproduktionens location, there is an adaptation of products to match the needs of linguistic and cultural, as well as the wishes of consumers.

Also there is a need for the contents production's localization that is an adaptation of the products to correspond to linguistic and cultural needs as well as the consumers' wishes.

Also we need the contents production's location, there is an adaptation of products to match the needs of linguistic and cultural, as well as the wishes of consumers.

Also we need the contents production's location, there is an adaptation of products to match the needs of linguistic and cultural, as well as the wishes of consumers.

Also the contents production's location is needed, that is an adaptation of products to match the needs of linguistic and cultural, as well as the wishes of consumers.

:	T1	T2	T3	T4	T5	N1	N2	N3	Ranks assigned:				
									1	2	3	4	5
Sentence 1													
	2	3	1	1	1	2	2	1	4	3	1	0	0
	3	2	1	3	2	3	3	2	1	3	4	0	0
	3	4	3	4	3	4	3	4	0	0	4	4	0
	3	4	3	4	3	4	3	4	0	0	4	4	0
	4	4	1	4	4	5	4	3	1	0	1	5	1
Sentence 2													
	1	2	1	1	2	2	2	2	3	5	0	0	0
	3	3	2	2	4	3	4	2	0	3	3	2	0
	1	4	4	4	3	4	3	4	1	0	2	5	0
	1	4	4	4	3	4	3	4	1	0	2	5	0
	2	5	4	4	5	3	3	3	0	1	3	2	2
Sentence 3													
	4	4	3	4	3	3	5	3	0	0	4	3	1
	3	2	4	3	3	5	5	2	0	2	3	1	2
	4	3	3	4	4	3	4	4	0	0	3	5	0
	4	3	3	4	4	3	4	4	0	0	3	5	0
	4	4	2	4	3	4	4	4	0	1	1	6	0

Sentence 4													
	3	2	1	2	2	2	3	1	2	4	2	0	0
	3	3	3	3	4	4	3	3	0	0	6	2	0
	3	3	4	3	4	3	4	3	0	0	5	3	0
	3	3	4	3	4	3	4	3	0	0	5	3	0
	3	4	4	4	3	3	4	3	0	0	4	4	0
Sentence 5													
	1	1	1	2	1	1	1	1	7	1	0	0	0
	3	4	1	2	3	3	2	2	1	3	3	1	0
	2	2	3	3	2	2	2	2	0	6	2	0	0
	1	2	1	3	2	2	2	2	2	5	1	0	0
	4	4	2	3	4	2	3	2	0	3	2	3	0
Sentence 6													
	2	2	1	2	2	2	2	2	1	7	0	0	0
	3	4	1	2	2	3	3	3	1	2	4	1	0
	4	4	4	3	4	4	4	5	0	0	1	6	1
	4	4	4	3	4	4	4	5	0	0	1	6	1
	2	2	2	2	2	3	3	2	0	6	2	0	0

Sentence 7														
	2	2	1	2	2	2	3	1	2	5	1	0	0
	4	4	5	4	4	4	4	2	0	1	0	6	1
	3	3	3	3	3	3	3	4	0	0	7	1	0
	3	3	3	3	3	3	3	4	0	0	7	1	0
	4	5	4	4	5	5	4	1	1	0	0	4	3
Sentence 8														
	2	2	1	3	2	2	2	3	1	5	2	0	0
	2	3	1	1	2	3	2	2	2	4	2	0	0
	3	4	4	4	4	4	3	5	0	0	2	5	1
	3	4	4	4	4	4	3	5	0	0	2	5	1
	2	3	3	3	4	5	4	3	0	1	4	2	1

Sentence 9													
	3	4	2	3	4	4	4	4					
									0	1	2	5	0
	2	3	1	2	4	2	3	2					
									1	4	2	1	0
	3	4	2	3	3	4	4	2					
									0	2	3	3	0
	2	2	1	3	3	2	3	1					
									2	3	3	0	0
	3	4	2	4	3	4	3	2					
									0	2	3	3	0

Sentence 10													
	3	1	1	3	2	2	3	1	3	2	3	0	0
	3	3	3	3	3	3	4	2	0	1	6	1	0
	4	3	3	3	4	4	4	4	0	0	3	5	0
	3	2	4	4	2	2	3	3	0	3	3	2	0
	3	4	5	5	5	3	3	4	0	0	3	2	3
Sentence 11													
	2	1	1	2	2	2	3	2	2	5	1	0	0
	5	4	5	4	3	4	3	3	0	0	3	3	2
	2	3	1	2	2	3	3	4	1	3	3	1	0
	4	3	2	3	4	3	3	3	0	1	5	2	0
	2	2	1	2	3	3	3	3	1	3	4	0	0
Sentence 12													
	2	4	2	3	4	4	4	3	0	2	2	4	0
	3	4	1	4	3	4	4	4	1	0	2	5	0
	3	4	2	4	4	4	4	4	0	1	1	6	0
	3	4	2	4	4	4	4	4	0	1	1	6	0
	3	4	2	5	5	4	4	4	0	1	1	4	2

Sentence 13														
	2	1	1	3	1	2	2	1	4	3	1	0	0
	3	3	2	3	2	2	3	2	0	4	4	0	0
	3	3	2	3	4	3	4	3	0	1	5	2	0
	3	3	2	3	4	3	4	3	0	1	5	2	0
	3	3	3	4	3	3	3	2	0	1	6	1	0
Sentence 14													
	2	3	1	3	2	3	3	2	1	3	4	0	0
	2	3	1	2	1	2	2	1	3	4	1	0	0
	2	3	1	4	2	3	3	2	1	3	3	1	0
	2	3	1	4	2	3	3	2	1	3	3	1	0
	4	4	4	5	5	5	4	5	0	0	0	4	4
Sentence 15													
	2	1	1	2	1	1	2	1	5	3	0	0	0
	5	4	2	3	3	2	2	2	0	4	2	1	1
	3	4	2	4	4	2	2	2	0	4	1	3	0
	3	4	2	4	4	2	2	2	0	4	1	3	0
	4	4	4	5	5	3	2	3	0	1	2	3	2
Sentence 16													
	2	1	1	1	1	2	2	1	5	3	0	0	0
	2	3	4	3	4	3	3	2	0	2	4	2	0
	2	2	2	2	3	3	2	2	0	6	2	0	0
	2	3	2	4	4	3	2	2	0	4	2	2	0
	2	2	2	3	4	3	2	2	0	5	2	1	0

Sentence 17													
	2	3	2	3	3	3	3	2	0	3	5	0	0
	2	2	1	2	3	5	2	3	1	4	2	0	1
	2	2	2	3	4	3	4	2	0	4	2	2	0
	2	2	2	3	4	3	4	2	0	4	2	2	0
	4	2	2	2	3	4	3	4	0	3	2	3	0
Sentence 18													
	2	1	1	1	1	2	2	1	5	3	0	0	0
	2	3	1	2	2	3	2	1	2	4	2	0	0
	4	4	4	3	3	3	3	3	0	0	5	3	0
	5	4	3	5	4	3	3	2	0	1	3	2	2
	5	5	3	4	5	3	3	2	0	1	3	1	3
Sentence 19													
	2	2	1	2	2	3	3	1	2	4	2	0	0
	3	2	3	3	2	3	2	1	1	3	4	0	0
	3	3	2	3	5	3	4	3	0	1	5	1	1
	3	3	2	3	5	4	4	3	0	1	4	2	1
	2	2	2	4	4	4	4	2	0	4	0	4	0

Sentence 20														
	1	1	1	1	1	2	2	1	6	2	0	0	0
	3	3	4	3	4	4	3	5	0	0	4	3	1
	3	3	2	4	2	3	4	3	0	2	4	2	0
	3	3	2	4	2	3	4	3	0	2	4	2	0
	4	3	2	4	3	3	4	2	0	2	3	3	0

Results from manual evaluation

	1	2	3	4	5	Check sum
SMT	53	64	30	12	1	160
RBMT	14	48	61	29	8	160
Lexical sub	3	33	63	58	3	160
Subtree sub	6	33	61	55	5	160
TED-R	3	35	46	55	21	160