Introducing corpus-based rules and algorithms in a rule-based machine translation system

Loïc Dugast

Doctor of Philosophy
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh
2013

Abstract

Machine translation offers the challenge of automatically translating a text from one natural language into another. Statistical methods - originating from the field of information theory - have shown to be a major breakthrough in the field of machine translation. Prior to this paradigm, many systems had been developped following a rule-based approach. This denotes a system based on a linguistic description of the languages involved and of how translation occurs in the mind of the (human) translator.

Statistical models on the contrary use empirical means and may work with very little linguistic hypothesis on language and translation as performed by humans. This had implications for rule-based translation systems, in terms of software architecture and the nature of the rules, which were manually input and lack any statistical feature.

In the view of such diverging paradigms, we can imagine trying to combine both in a hybrid system. In the present work, we start by examining the state-of-the-art of both rule-based and statistical systems. We restrict the rule-based approach to transfer-based systems. We compare rule-based and statistical paradigms in terms of global translation quality and give a qualitative analysis of their respective specific errors. We also introduce initial black-box hybrid models that confirm there is an expected gain in combining the two approaches.

Motivated by the qualitative analysis, we focus our study and experiments on lexical phrasal rules. We propose a setup allowing to extract such resources from corpora. Going one step further in the integration of rule-based and statistical approaches, we then examine how to combine the extracted rules with decoding modules that will allow for a corpus-based handling of ambiguity. This then leads to the final delivery of this work: a rule-based system for which we can learn non-deterministic rules from corpora, and whose decoder can be optimised on a tuning set in the same domain.

Acknowledgements

This has been a long journey. With all kinds of weather, sometimes all four seasons in one day. There are places like that, where you just have to get by with ever-changing skies. With the right attitude though, you should be able to appreciate the landscape. And the people you meet along the road. Many of them have been of great help.

I would like to especially thank my supervisor Philipp Koehn for his continuing support.

Along the road and under the same latitude, I have to mention my second supervisor Miles Osborne and the whole Machine Translation crew in Edinburgh, with special stars for my office mates Abhishek Arun and Hieu Hoang.

I remember how Markus Becker and Carsten Brockman answered my questions late at night in the office when I had just arrived. Thank you too Amittai Axelrod, Alexandra Birch, Phil Blunsom, Trevor Cohn, Adam Lopez... I must be forgetting many of you, please do not take offence.

On the other side of the Channel, thank you to Jean Senellart of Systran.

Last but not least, I cannot spare my parents from being mentioned in this section. Thank you for being here, for encouraging me in those foreign lands of studies, again, and without judging me.

Well thank you all family and friends from Scotland or France, for both bearing with me and supporting me.

Voilà.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Loïc Dugast*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

If this thesis was about history of science, I would begin this introduction in such a way: Once upon a time, as computers were about to be invented, scientists started to daydream about the possibilities they would open for humanity. Quickly, the fable of a non-human intelligence was being embroidered. Some would think in terms of mathematics, replacing human problems with equations and letting computers, well... compute the solutions. Some others would dream of feeding machines with vast amounts of human knowledge, and teach them how to process it, letting computers mimick human reasoning and make it much quicker.

To give the big picture of what is at stake here, we may indeed trace the dichotomy beyween the two above-mentioned approaches back to the early days of Computer Science. It arose between the newly founded Information Theory by Shannon, which builds mathematical models of information as a whole, and the birth of higher level programming languages, which was born at the same time as formal linguistics, propelled by the major work of Noam Chomsky. With programming languages came the need to distinguish between lexicon, syntax and semantics. All of this had to be described mathematically so that computers could process the different steps involved. Both descriptions became essential in the rise of formal linguistics and this new discipline called *Natural Language Processing (NLP)*. As much as information theory came from the hardware side of computer science and was used to deal with the lower layers of human linguistic processing, i.e. phonetics, the theory of formal languages dealt with advanced software processing and was from early on connected with attempts to model the syntax of natural languages.

Soon in the history of computer science, people would think of hard intellectual tasks that humans perform, that machines could be designed to achieve, possibly even better. Among them was the task of translating a text from one language to another. Machine Translation, that is. Very early, pioneers of computer science would struggle with house-sized machines that required cardboard to be punched to let instructions in. Such machines could only process little data at a time, which made the data-based approach less relevant in practice. All of this was an inducement to see machine translation as the implementation of modern advanced formal theories of grammars. At the same time it was made clear very quickly (Chomsky, 1957) that formal languages sufficient to encode *programming* languages were short of describing far more complex *natural* languages. Machine translation systems that required linguists or lexicographers to enter rules and dictionaries quickly departed from the algorithmically sound formal languages.

Meanwhile, it took machines to reach a certain level of computing power for the first mentioned approach to gain ground, as actual implementation and experimentation of those ideas became possible. That is, machine learning instead of knowledge-based artificial intelligence.

Yet, in the past twenty years where such empirical systems dubbed *Statistical Machine Translation* systems took momentum and started to outperform complex, *Rule Based Machine Translation* systems, difficult language pairs (in terms of syntactic differences or even monolingual complexities in morphology) still resisted.

With the rise of statistical approaches, the use of automatic metrics also came into play, based on a simple string-matching counting that uses one or more reference translations of a given text. The use of such a proxy to evaluate a machine translation output is supported by studies which show a correlation with human judgements. Yet it was shown (Callison-Burch et al., 2006) that this correlation is much lower in the case of rule based than in the case of statistical systems.

The latter grounds and the difficulty to incorporate linguistic knowledge in formalised statistical models motivate the need for a study on how both approaches can be combined in an effective way. This is what we intend to do in the work we present here.

1.2 Overview

1.2.1 Contribution of the thesis

The work presented initiated in the shift in research from rule-based to statistical machine translation. All this while the former approach still yielded better results according to manual evaluation. This led to an initial hybrid system, combining a rule-based and a statistical phrase-based system (Dugast et al., 2007), and later on to a statistical model of a rule-based system (Dugast et al., 2008). Dugast et al. (2009a) explored the possibility of augmenting a rule-based system thanks to statistical approaches.

This thesis contributed to these three following aspects:

- Rule-based and statistical systems produce different types of errors. We produced a set of systems from the same dataset, compared translation quality of all, reproducing prior results on a lower correlation between automatic score and manual evaluation of translation quality for rule-based systems (Callison-Burch et al., 2007). Yet the main contribution of this study lies in the error analysis which shows complementarities between the rule-based and the statistical approach.

- Rule-based and statistical systems can be combined to produce better results over both baselines. Hybrid systems were submitted in an international machine translation workshop, which managed to fare better than its purely statistical competitors (Dugast et al., 2007). The *Statistical Post Editing* setup has been a commercial reality for the past few years.

- Lexical coverage is the main type of improvement expected from empirical methods when used to augment a rule-based system. The dictionary extraction setup has also been a commercial reality for the past few years.

1.2.2 Structure of the document

In Chapter 2, we describe a specific rule-based system and a statistical system. Following this, we present preliminary experiments with black-box combinations of those systems.

Chapter 3 presents experiments on the systems described in the previous chapter. Hybrid systems in particular provide significant improvements over both the rule-based and the statistical, as shown quantitatively by automatic evaluation scores and qualitatively thanks to a manual analysis.

In Chapter 4, we present results on the extraction of lexical rules for the rule-based system. They are first used to provide a unique translation option for each source phrase. In that case a discriminative selection of these entries is shown to ensure an improvement over the baseline. In addition, we provide oracle results which show that lexical ambiguities alone offer a wide margin for progress in the rule-based setup.

Chapter 5 presents a setup where both extracted entries and the base entries in the rule-based system can be combined with a statistical disambiguation module and get better results.

1.2.3 Publications

In (Dugast et al., 2007), we presented initial results on a black-box combination of a rule-based system with statistical decoding. We also provided a qualitative analysis of the changes brought by the statistical layer. In the later work of Ueffing et al. (2008), enhancements were brought to the initial post-editing model.

(Dugast et al., 2008) is an attempt of using the rule-based system to produce a synthetic corpus. It shows that the rule-based translation performance can be reproduced, at least in terms of an automatic translation quality metric, using source and target language monolingual corpora.

We presented experiments on the extraction of lexical rules in (Dugast et al., 2009b), while (Dugast et al., 2009a) showed that the extracted rules could efficiently improve the existing rule-based system.

Chapter 2

Literature review

Machine Translation (MT), although currently dominated by Statistical Machine Translation (SMT) still see the newest systems coexist with running commercial Rule-Based Machine Translation (RBMT). For most working in this field of research, it is thus unclear what RBMT exactly means. Moreover, beside of advanced technical discussions and breakthroughs on algorithms, language-specific issues, the same fundamental questions keep being risen such as how to evaluate the success of an MT task. In this chapter, we try to disentangle issues of definition and evaluation of success, before looking into related work, mainly within the realm of system combination and hybrid approaches.

This chapter intends to give an overview of machine translation, especially both rule-based and statistical. It then gives arguments for working on hybrid systems. Finally, related work is reviewed.

2.1 Machine Translation

Translation between languages involves the composition of a text in a (*target*) language from the meaning of an existing text in another (*source*) language. It is performed by a speaker with a knowledge of both languages, generally a native speaker of the *target* language, who from the understanding of the source text, produces a target text with the equivalent meaning. As a human task it is therefore driven by two objectives, often contradictory: to purport the meaning (*fidelity*) and to produce a text that is fluent in the target language (*transparency*).

One of the earliest challenges imagined in Artificial Intelligence (AI), Machine Translation (MT) aims at automatizing translations. It has to respond to the two goals above mentioned, which are rather called *adequacy* and *fluency* in the MT literature.

Why is this a difficult problem? Well, in addition to the tension between adequacy and fluency, the sole problem of understanding the meaning of the source text involves solving ambiguities at many levels: lexical, syntactical, semantical and even pragmatical. In fact, one could argue that this would require to solve many problems in Natural Language Processing which are hard to model. In addition, current models are computationally expensive.

As for other problems in AI, one first attempted to solve it through a reproduction of what was intuitively understood of the mechanisms involved in manual translation. This meant first constructing a semantic representation of the source text, before generating a projection of this meaning in the target language. Linguistic theories on syntax and lexicon, if not in semantic analysis were there to help.

Alternatively, pattern recognition and machine learning methods had already started to develop, though limited by computational means. Solving a problem in that framework did not make any assumption on the hypothetical “real” linguistic process of translation, but instead relied on “passing the Turing test”. We could be content with the machine *simulating* the work of a human being, independently of what actually occured in the process of a human translation. We will see further how these two approaches result in different implementations. They are not necessarily contradictory however and we see how their convergence is of interest.

2.2 Rule-Based Machine Translation

> *I took some texts in Russian and figured out a scheme for transliterating the Cyrillic characters so that I could input them and experiment with translation into English. Before long I had worked out several algorithms, and I began to produce translations.* Peter Toma, founder of Systran

It is hard to define what exactly the term ”rule-based system” means. Even for the same language pair, a number of very different machine translation systems have been developed based on manually entered rules, starting with the most simple tools human translators use: bilingual dictionaries. Beyond this common aspect, we see a number of different design choices and implementations.

In this section, we try to define the core properties of a rule-base system, regardless of less important design choices and implementation details. Though we will have to experiment with a specific implementation, we try here to retain the most general aspects.

Initial practical ideas in Machine Translation mostly reflect, albeit simplified the work of a human translator. Hence the famous Vauquois triangle (Figure 2.1), that illustrates how the translator reads a source text, understands it (thus forms a mental picture of it, an interlingua representation) and then, going down the slope and using his competence (Chomsky, 1965) in the target language, creates a target text by making syntactic and lexical choices which adequately project the *meaning*. Such a view is naturally unlikely to be implemented, for the sole reason that this would require solving the problem of natural language understanding itself. Consequently, early implementations of rule-based systems lowered the bar and tried to instead produce intermediate representations. They introduced transfer rules between them, starting with bilingual dictionary entries.

We find in the literature attempts of ”dictionary-based” machine translation (at the lower end of the scale) and ”interlingua” machine translation (at the other end, claiming to use a higher level language-independent semantic representation). Yet, most systems, including currently used commercial rule-based systems are located in the middle of this scale: transfer-based machine translation. In this thesis, we choose to restrict the scope of *rule-based* systems to *transfer-based* systems. Such implementations rely on translation rules at an intermediate level between surface forms and the ideal interlingua. In the following pages, the term *rule-based system* will be used in this narrow sense.

2.2.1 A few examples of rule-based systems

Starting from bilingual dictionaries and trying to implement advances in formal linguistics theories, a few systems have reached a stage sufficient for practical usage, at least in some applications. Among such systems that are still maintained and in-use, the SYSTRAN system (Toma, 1972) may be one of the earliest. According to the publicly available descriptions, the PROMT system and the Logos system are similar. However, such systems have reached different levels of development for different languages. The METAL project (White, 1985) ended up in a few commercial applications. This feature-rich transfer-based system, running on a LISP machine used constraint programming that could handle some level of ambiguity in the process. More recently, the Apertium project (Ramırez-Sánchez et al., 2006) implements a shallow transfer based on manually written rules with dictionary entries.

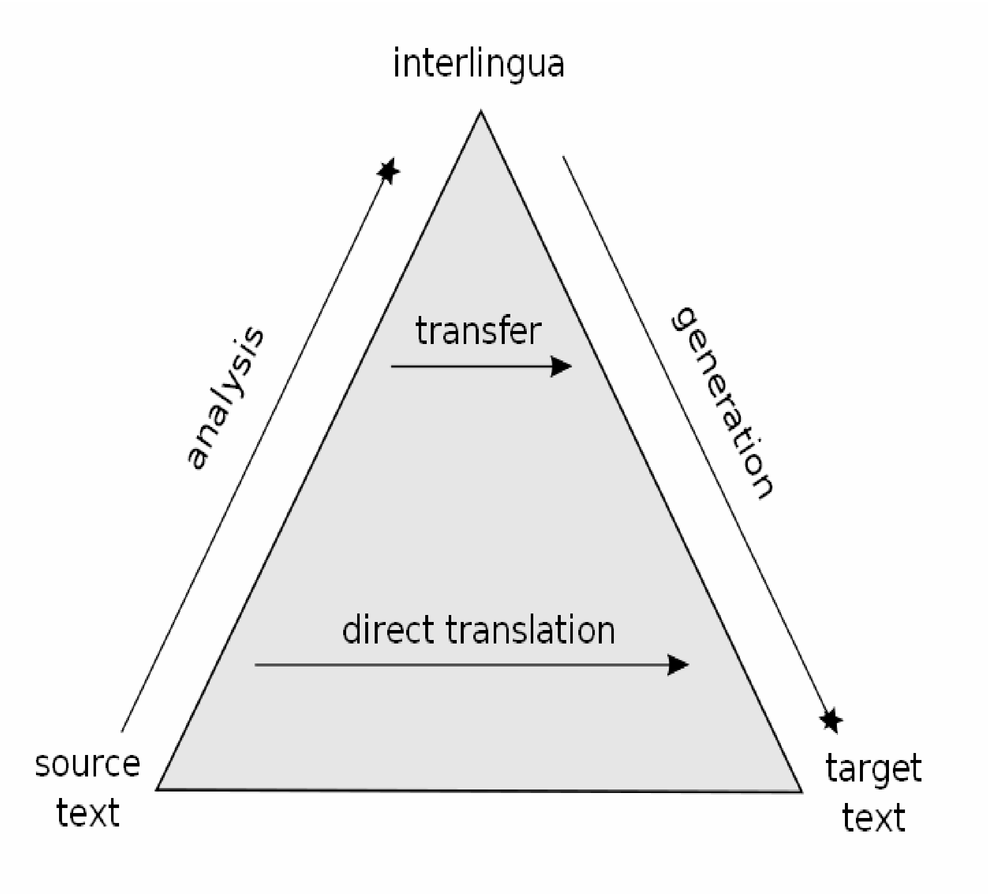


Figure 2.1: The Machine Translation triangle

2.2.2 A generic rule-based system

In this section, we aim at presenting generic features of our definition of a rule-based system. A first characteristic we retain for our definition of a rule-based system is the splitting of the translation process between source analysis, transfer and synthesis of the target sentence. The analyzed source sentence may be described as a dependency or a syntactic tree. Transfer rules may come in various types. They may be very general, structural rules, such as to translate a noun phrase from French to English, where leaves are translated by dictionary entries such as: bleu → blue, chat → cat. Yet, disambiguation rules may be specified in the form of decision trees, for example from English to French: bank (if POS=noun AND ”money” in context) → banque, bank (if POS=noun AND ”river” in context) → rive.

The translation process happens in sequence: ambiguities are solved at each consecutive level, without considering globally the combinations of these choices. For instance, part-of-speech is disambiguated first, then a parse tree of the input is decided on, then transfer rules are chosen (they are either unambiguous or use disambiguation routines when available, as mentioned above). Finally, choices are made to generate the target sentence.

In other words, although there might be a few counter-examples in the existing systems, we assume in this work that a rule-based system does not handle ambiguity globally.

Rules are entered manually and are therefore also motivated by some human understanding of the language, making them readable and editable.

As a consequence of both the required effort to enter rules and the level of generalization of these rules, the construction of the rule set is incremental and the number of rules remains small as compared with statistical systems.

Let us sum up the features that define a rule-based system:

1. it follows a source-transfer-synthesis process
2. ambiguities are solved sequentially
3. rules are entered and understandable by humans
4. rules display generalization (always beyond surface words, at least using morphological generalization)
5. rules are added incrementally and their total number remains small

2.2.3 Example: the SYSTRAN system

We describe here the SYSTRAN system, in terms of both architecture, algorithms, and manually built resources. What is specific to SYSTRAN with respect to the prototypic rule-based system we have just defined ?

- analysis is dependency-based
- a dictionary coding engine is available
- language pairs have received various amounts of efforts: total of dictionary entries ranges from 50k to half a million
- some domain adaptation is possible thanks to domain-specific dictionaries

SYSTRAN's first prototype was built in 1968 to translate from Russian to English. It currently includes translation engines for 80 language pairs covering 22 source languages. Numerous years of development making use of various techniques make it difficult to classify. Yet, the best approximation would be to consider it as a transfer-based system making extensive use of large dictionaries, both monolingual and bilingual. We try here to describe the original rule-based system, regardless of the current developments which tend to incorporate corpus-based methods.

As can be seen on Figure 2.2, translation starts with some preprocessing including document filtering (aiming at separating text from any other kind of data), plus segmentation into paragraphs and sentences. Different dictionary look-ups are performed sequentially. The first dictionary to be looked up is the ”idiom dictionary”, which contains idiomatic sentences or phrases. A dictionary of (single-word) stems or ”main dictionary” is then consulted, before a dictionary of phrasal entries called ”Limited Semantics Dictionary”. These look-ups do not involve the use of any grammatical analysis, but are instead the first step preparing for it.

Analysis then tries to solve part-of-speech ambiguities, before phrases and then clauses are identified. At that stage, linguistic routines are then performed to construct syntactic dependencies and finally, an analysis tree. Figure 2.3 gives an example of the SYSTRAN analysis. The transfer phase starts by looking up the ”Conditional Limited Semantics Dictionary” or ”CLS” which is more or less a dictionary of lexical disambiguation rules. More lexical routines are then performed before the synthesis phase can start. In this last stage, remaining words (not translated by the CLS dictionary) are translated and inflected, and syntactic rearrangement is performed to fit the target language.

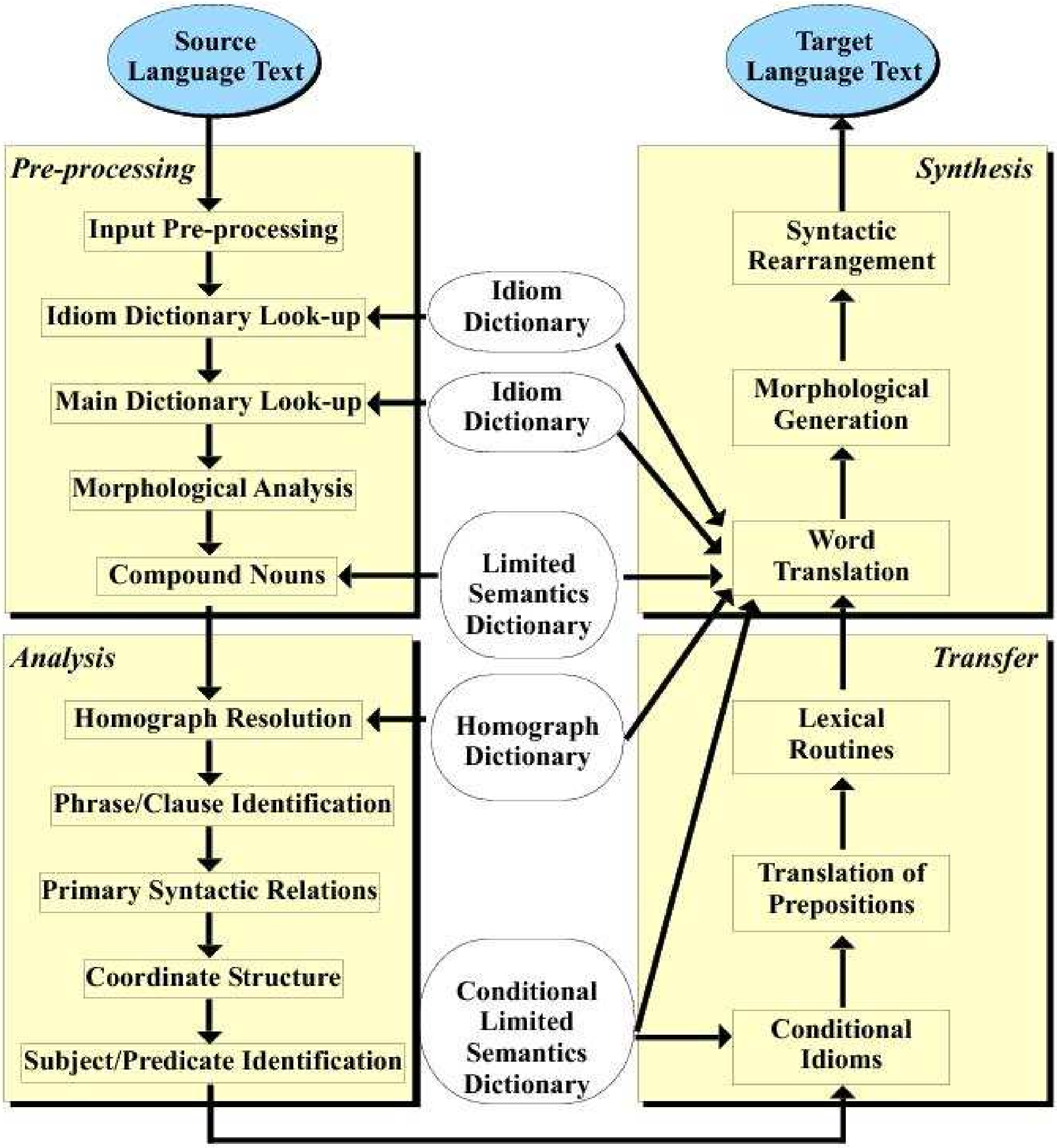
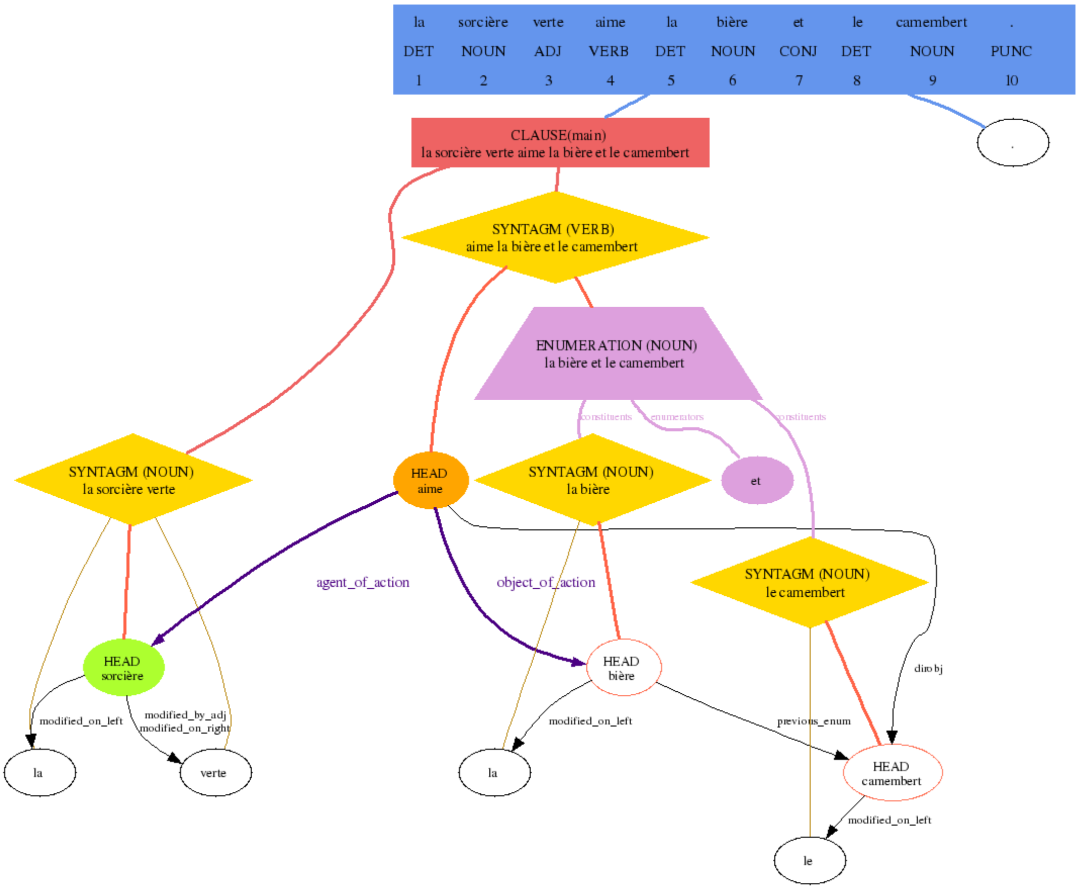


Figure 2.2: SYSTRAN translation process (from Alex (2002))



la	sorcière	verte	aime	la	bière	et	le	camembert	.
DET	NOUN	ADJ	VERB	DET	NOUN	CONJ	DET	NOUN	PUNC
1	2	3	4	5	6	7	8	9	10

Figure 2.3: SYSTRAN analysis example

The greedy dependency parser creates head-modifier links and identifies subject and object of predicate.

2.2.4 A description based on the linguist's understanding of the process

We aim here at clarifying the nature of the rules used in the rule-based system. First of all, the whole translation problem is subdivided into sub-problems: source analysis, transfer and final generation of the target sentence. This framework follows a linguistic description of the translation process, *related to contrastive linguistics*. In a constrastive grammar such as proposed by Salkoff (1999), each phenomenon encountered in the human translation process is reduced to a few formal rules.

The actual rules have then to be entered within this framework of sequential decisions.

Each rule belongs to one of the stages which each takes a unique decision on the ambiguity they are meant to solve. This decision is passed to the following module in the workflow to process the next type of ambiguity.

The whole translation sequence is examplified in Table 2.1.

2.2.4.1 Source analysis

Analysis of the source sentence starts with the tokenization stage, which is driven by the dictionary entries. The segmentation ambiguity that results from the dictionaryentries is managed by a plain default behaviour: longer entries overule shorter ones and in the case of overlapping entries of same length, either the right-hand side or left-hand side one is chosen for a given source language.

The next step is Part-Of-Speech disambiguation. Part-of-Speech disambiguation rules are coded in the form of manually entered consecutive steps. Similar to the Brill tagger (Brill, 1992), it assigns an initial default morphological tag to each token. Then correction rules are applied sequentially to modify this initial stage.

The delimitation of clause boundaries is the first parsing step. Again, this uses a sequence of processing steps and uses very general lexical anchors to identify main and subordinate clauses. It may typically fail in the case of non-detected embedded clauses (*”The outcome of the negotiations* ***if they succeed*** *might be surprising.*”) or lack of lexical anchors (*such as the absence of a conjunction introducing the clause in ”I am pretty sure this is a Glenfiddich.*”).

In the next step, local syntactic relationship are identified based on the features attached to the matched source words (subcategorization,transitivity of verbs).

The main subject and predicate are then identified in each clause.

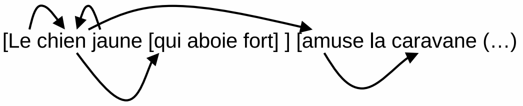
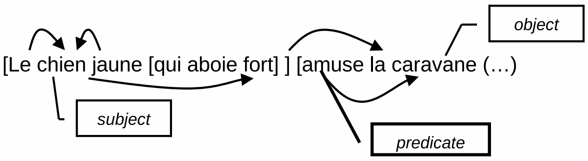
1	Source text	(…) prochaine oasis. Le chien jaune qui aboie fort amuse la caravane, l'émir et en particulier son chameau; ce dernier blatère bruyamment. Les jours (…)
2	Source sentence	Le chien jaune qui aboie fort amuse la caravane, l'émir et en particulier son chameau
3	Basic tokenization and normalization	Le chien jaune qui aboie fort amuse la caravane , le émir et en particulier son chameau
4	Dictionary matching	Le [chien] [jaune]qui [aboie] [fort] [amuse] la [caravane] , le [émir] et [en particulier] son [chameau]
5	Part-Of-Speech disambiguation	Le [chien]*(noun)* [*jaune](adjective) qui* [aboie]*(verb)* [fort] *(adverb)* [amuse] *(verb)* la [caravane] *(noun)* , le [émir] *(noun)* et [en particulier] *(adverb)* son [chameau] *(noun)*
6	Clause boundaries	[Le chien jaune qui aboie fort] [amuse la caravane , le émir et en particulier son chameau]
7	Recursive search for embedded clauses	[Le chien jaune [qui aboie fort]] [amuse la caravane , le émir et en particulier son chameau]
8	Syntactical relationships	
9	Role identification	 object [Le chien jaune [qui aboie fort]] [amuse la caravane (…) subject predicate
10	Intermediate semantic representation	(here, identical with the previous step)
11	Lexical transfer	chien -> dog ; aboyer -> to bark ; …
12	Rearrangement	([dog yellow] -> [yellow dog]) yellow dog who bark loud amuse caravan
13	Insertion of prepositions, determiners	the yellow dog who bark loud amuse the caravan
14	Inflection	the yellow dog who barks loud amuses the caravan
15	Detokenization and casing	The yellow dog who barks loud amuses the caravan, the emir and especially his camel

Table 2.1: The translation sequence

2.2.4.2 Intermediate meaning representation

A last transformation aims at normalizing equivalent syntactic forms (such as active and passive in languages for which this is relevant) into one semantic dependency representation.

2.2.4.3 Transfer

Transfer rules consist mainly in *lexical* transfer rules. Meanings are retrieved from the bilingual dictionaries. First, disambiguation rules are applied as a decision list where each rule is based on immediate context and dependency analysis. Figure 2.4 gives an example of such a rule. On this figure, the *EXIT* final node has been duplicated to ease reading. The *Input* is the French word whose inflected form matches the verb ”devoir”. On a side note, although this very example remains readable, some of the rules have become increasingly complex due to constant modifications of the system. Also, rules that were initially motivated by linguistic reasons have been changed by add-hoc modifications. Which might be another reason for a decreasing readability of some of the rules overtime. For the remaining source words, default translations are applied. Expressions which require a structural modification of the dependency graph are also handled at this stage.

Note that such rules would be difficult to learn automatically because of both the numerous features used and the small number of occurences which may be found in a corpus.

2.2.4.4 Target generation

The generation stage takes the previously mentioned meaning representation as input. In order to produce a sentence in the target language, it is necessary to produce a compatible inflection of the target words which both translates the source inflection and respects agreement where needed. Grammatical words such as prepositions and determiners have to be inserted according to the target language generation rules. Additionally, rearrangement rules aim at producing a correct word order according to the target language specificities.

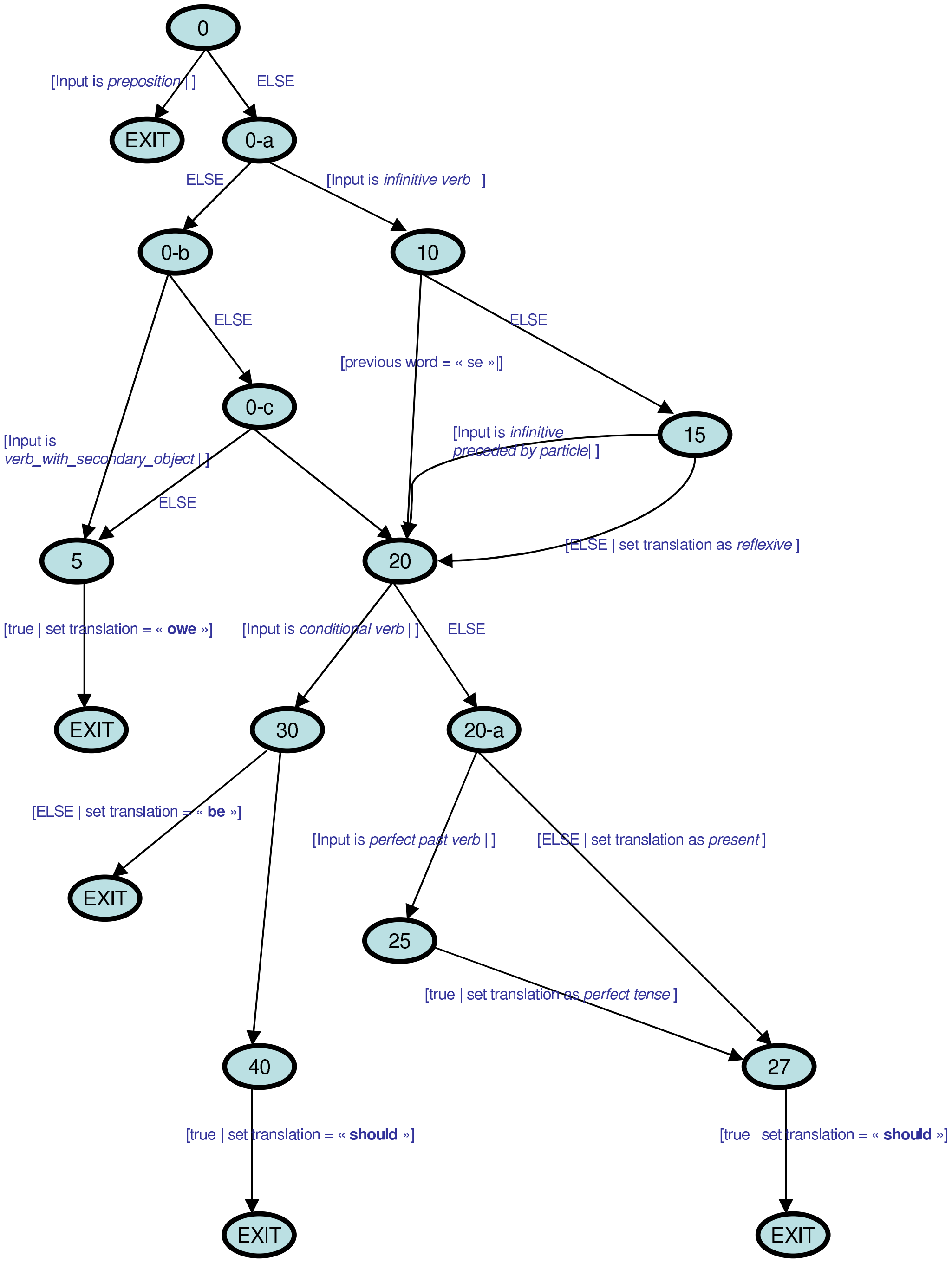

0
[Input is *preposition* |]
ELSE
EXIT
0-a
ELSE
[Input is *infinitive verb* |]
0-b
10
ELSE
ELSE
[previous word = « se »|]
0-c
[Input is *infinitive preceded by particle*|]
15
[Input is *verb_with_secondary_object* |]
ELSE
5
20
[ELSE | set translation as *reflexive*]
[true | set translation = « **owe** »]
[Input is *conditional verb* |]
ELSE
EXIT
30
20-a
[ELSE | set translation = « **be** »]
[Input is *perfect past verb* |]
[ELSE | set translation as *present*]
EXIT
25
[true | set translation as *perfect tense*]
40
27
[true | set translation = « **should** »]
[true | set translation = « **should** »]
EXIT
EXIT

Figure 2.4: Example transfer rule: ”devoir” French verb to its English translation

2.2.5 How lexicographers enter new rules

Lexicographers spot errors in the translation output and interpret which phenomenon may be incorrectly described or simply not covered by the current set of rules. They have to identify whether this is an analysis, transfer or synthesis error, or a lexical error. They then come up with rule(s) that would fix the observed error. They test it on a few sample source sentences to compare it with the baseline. The corpus used for testing is of a wide coverage domain, made up of various sources, mostly news. A threshold of improved over degraded translations then serves as a selection criterion to accept the new rule. Again, a major difficulty for modifying some of the stages (such as stages involved in the dependency analysis) lies in the increasing complexity they display, as rules are added and get more sophisticated. Overall, the most frequently modified or augmented set of rules is by large the database of bilingual dictionaries. As for any type of rules, lexicographers have to cope with the surrounding framework. First of all, lexical rules are not hierarchical, which imposes constraints on the use of multiword phrasal units. Word segmentation (the initial phrasal segmentation) is induced by dictionary matches, such contiguous entries enforce a contiguous target phrase. Another issue lies in side-effects in the source analysis. Yet another aspect lies in the choice of a single translation for each source phrase: if no disambiguation criteria are provided, it has to be chosen as the most ambiguous possible (so that adequacy is preserved most of the time). Otherwise, there is a possibility of attaching disambiguation clues to the entry (conditions on context words, possibly using the labeled edges of the dependency analysis). Still, in practice, a large majority of entries do not carry any such clue. As a conclusion, if the sequential processing of the translation process is aimed at enabling the use of simpler rules that are more manageable by humans, spotting and understanding undesired interactions between the various rules constitutes in return a major difficulty for the lexicographer (Bod, 1992).

2.3 Statistical Machine Translation

Statistical techniques in Natural Language Processing range from simple word counting heuristics to advanced Machine Learning implementations. The human input in constructing a statistical translation model is far smaller than in the case of a rule-based system. This input comes in two kinds. Humans may be required to annotate a corpus explicitly (for instance, provide part-of-speech annotation) or implicitly (producing parallel corpora as a side-effect of a translation activity). The other kind of input deals with the architecture of the system, which implies both a training and a run-time framework. This goes from the implicit choice of rules through, for example, the crafting of a generative story of the translation process, down to the choice of a large number of features to be used by a Machine Learning algorithm to learn a classifier.

2.3.1 Models

> *One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ’This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.* Warren Weaver

2.3.1.1 The noisy channel model

In statistical machine translation we consider that all English sentences (e) are possible translations of a French (f) sentence, though with different probabilities. If the statistical model is good enough, better translations are given higher probabilities. We are thus looking for the highest probability translation(s).

$$\hat{e} = \arg\max_{e} P(e|f) \tag{2.1}$$

Bayes’ rule transposes the problem into finding the English sentence maximising the product of two components:

$$\hat{e} = \arg\max_{e} P(e) \cdot P(f|e) \tag{2.2}$$

The first component is the prior probability of seeing this English sentence. The second component is the conditional probability of having ’crypted’ (in the sense of Warren Weaver’s quote above) this English sentence into the French sentence we want to translate. This view of the ’noisy channel model’ (Figure 2.5) and was first presented by Brown et al. (1988) to solve the problem of machine translation.

Following the noisy channel model, we still have to explain a bit more the two main models involved. First, in order to model the probability of an English sentence, we have to break it up into a function of its components. We have to do that because we cannot just count every single English sentence: the generative property of human language makes data sparsity unavoidable even for the largest corpora we may have.

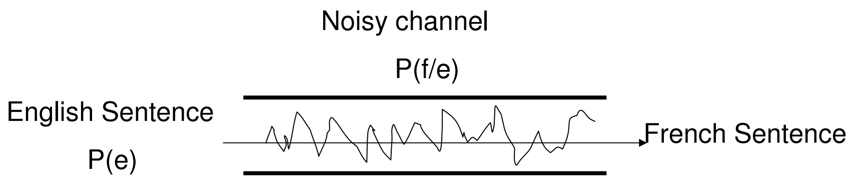


Figure 2.5: The noisy channel generative story

N-gram models constitute the most common and possibly most simple probabilistic language model. This model breaks up P(e) into conditional probabilities of a word to occur given its left-side context in the range of N-1 words. This is called an N-gram language model.

We shall not extend our review of language modelling but rather focus on the second component of the noisy channel model, i.e. translation modelling.

2.3.1.2 Word and phrase-based models

The most simple translation model one may think of consists in modelling the translation of the French sentence as the product of the translation of all its words separately. This constitutes the 'IBM model 1', that was presented in Brown et al. (1993). Four more such models of word-by-word translation of increased complexity are presented in the latter work.

In order to learn those probabilities, we are faced with an 'incomplete data' problem: if only we knew, for each sentence pair, which word is translated by which word, we could sum the counts and would immediately have our statistical model. Conversely, if only we had this statistical model, we could (through a decoding method, looking for the global optimal probability $P(f|e)$ or $P(e|f)$) for each sentence pair find the optimal 'alignment' between English and French words. This alignment problem (Figure 2.6) can be typically solved by the Expectation-Maximisation algorithm.

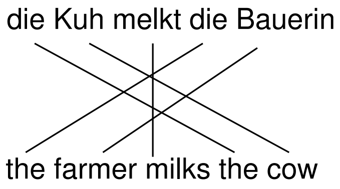


Figure 2.6: German-English word alignment

Word-alignment programs such as GIZA (Och and Ney, 2003) use a sequence of word-based translation models in sequence, the simpler model initializing the next, more complex one.

A major weakness however of word-based translation models is their not taking context words into account, which is not completely compensated by the use of n-gram language models when decoding. Using sequences of words instead is one way to carry context. These models are named ’phrase-based’ models (Koehn et al., 2003), though the ’phrases’ involved do not have to be linguistically motivated but are instead plain sequences of words. Figure 2.7 illustrates the way a sentence is translated in such a framework, usually using *beam search* decoding. This is currently the state of the art in statistical machine translation, though syntactic models are now starting to reach the performance of the phrase-based systems.

2.3.2 Training

Training is the off-line procedure which builds a statistical translation model out of corpora, that would be later used to search for an optimal translation of the source sentence.

2.3.2.1 Rule extraction

Although the ultimate goal is to maximize the translation quality (which can be approximated by an evaluation metric) on an unseen text, the extraction of rules is done independently from that objective function. It uses both statistical models (as word translation models) that are not directly related to the final task, and heuristics based on frequency counts. Following the idea of the noisy channel model, both a translation

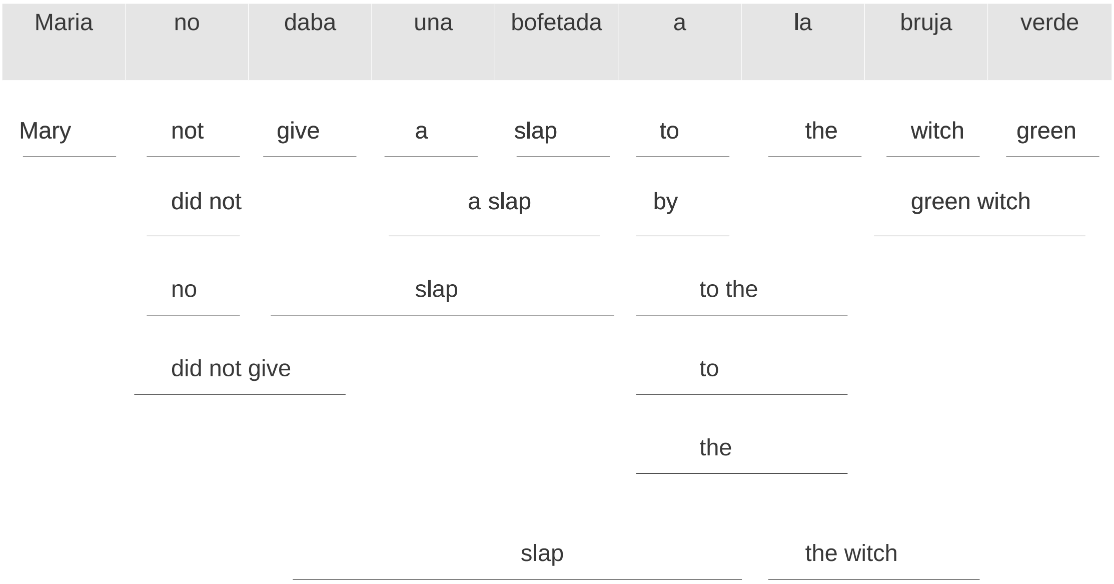


Figure 2.7: Translation ambiguity in a phrase-based framework (based on Koehn (2009))

Source	**Target**	**P(F|E)**	**P(E|F)**
chat et à	game	9.80E-03	1.00E+00
chat ne seront sûrement pas	cat are not likely to find that	1.00E+00	2.50E-01
chat ne seront sûrement pas	cat are not likely to find	1.00E+00 8	2.50E-01
chat ne seront sûrement pas	cat are not likely to	1.00E+00	2.50E-01
chat ne seront sûrement pas	cat are not likely	1.00E+00	2.50E-01
chat ne seront	cat are	1.00E+00	1.00E+00
chat ne	cat	3.33E-01	1.00E+00
chat	cat	6.67E-01	1.00E+00
chat échaudé craint l' eau	fool me once , shame	3.33E-01	3.33E-01
chat échaudé craint l' eau	fool me once ,	3.33E-01	3.33E-01
chat échaudé craint l' eau	fool me once	3.33E-01	3.33E-01
chat échaudé craint l'	fool me once , shame	3.33E-01	3.33E-01
chat échaudé craint l'	fool me once ,	3.33E-01	3.33E-01
chat échaudé craint l'	fool me once	3.33E-01	3.33E-01
chat échaudé craint	fool me once , shame	3.33E-01	3.33E-01
chat échaudé craint	fool me once ,	3.33E-01	3.33E-01
chat échaudé craint	fool me once	3.33E-01	3.33E-01

Table 2.2: Sample of a phrase-table for French-English

model and a target language model are used. In the vanilla phrase-based system, only sequences of words are considered, without any other linguistic structure than a notion of sentences (the parallel corpus is an alignment of *sentence pairs*). The translation model requires aligning subsequences of words, while the target language model may be learnt from the target side of this parallel corpus and/or any other target language text. Phrase alignment is done in two steps: an initial word alignment (Brown et al., 1993) and a phrase extraction heuristic. The output is a scored set of aligned sequences of source and target words (Table 2.2). Statistical features are attached to each phrase pair.

The training of the n-gram language model involves a basic counting of the frequency of each n-gram (we use *five* gram language models in the further experiments) with additional smoothing and discounting methods that lower the test-time perplexity.

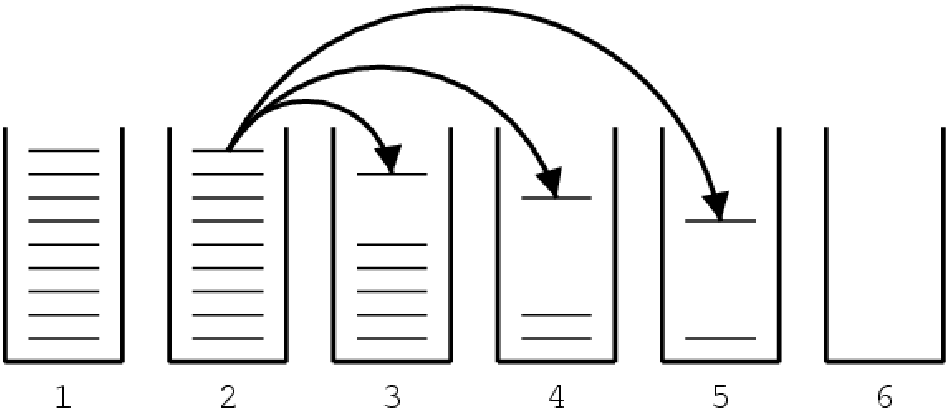


Figure 2.8: Filling in of stacks in beam search, as more foreign words get covered (graphic borrowed from Koehn (2009))

2.3.2.2 Discriminative tuning

In practice, a statistical translation model is often a combination of multiple features. A most common way of embedding the various submodels is log-linear combination. In a log-linear combination, each sum model contributes with the logarithm of the probability assigned to the specific feature. The total score assigned to a hypothesis is a weighted sum of all logarithms.

Thus, the use of a combination of statistical models in a log-linear framework (Equation 2.4) poses the need for a tuning of the relative weights given to those models. Given that these parameters are global, this is a possible means of bridging the gap between the non discriminative nature of rule extraction and the objective function of the evaluation metric that approximates translation quality on a held-out corpus. The algorithm used here is the Minimum Error Rate Training, an implementation of the Powell optimization algorithm (Och, 2003) that is based on the n-best list of translation outputs by the decoder.

$$\begin{aligned} \hat{e} &= argmax_e p\left(e_1^I | f_1^J\right) && (2.3) \\ &= argmax_e \sum \lambda_m h_m(e_1^I, f_1^J) && (2.4) \end{aligned}$$

2.3.3 Decoding

Decoding to find the optimal translation according to the model is complex (NP-complete, as shown by Knight (1999)) and hence unsolvable by exact search. We thus have to resort to dynamic programming algorithms with some pruning of the search space. The decoder we use implements a *beam-search*: the target sentence is generated from left to right, placing translation hypotheses in stacks (each defined by the number of foreign words covered, see Figure 2.8) which are pruned according to an estimation of the future cost (Koehn, 2009).

The use of a dynamic programming algorithm theoretically allows us to find an optimal solution in a large search space. And, although for matters of computability in a reasonable amount of time, pruning strategies are necessary, solutions found are very close to the optimal.

2.4 Hybrid rule-based and statistical MT

From a qualitative perspective, we may want to keep the following strengths of the rule-based systems (as evoked by Thurmair (2005)):

- the linguistically motivated rules of different levels allow for the description of complex phenomena
- keep the number of actual translation rules low so that they might be examined and edited manually
- allow for incremental learning of the model, in the fashion of dictionary editing in a rule-based system

On the other hand, there are missing features in the rule-based framework, which are core of the statistical one that may explain a sometimes lower performance than statistical systems while remaining compatible with its architecture:

- trainability from corpus (opposed to manually entered rules) which should enable easy domain adaptation (as opposed to rules in the RBMT system, which are often chosen to be as general as possible)
- tuning to a given metric (opposed to the evaluation of *improvement* by linguists, based on only a few samples because of time constraints)

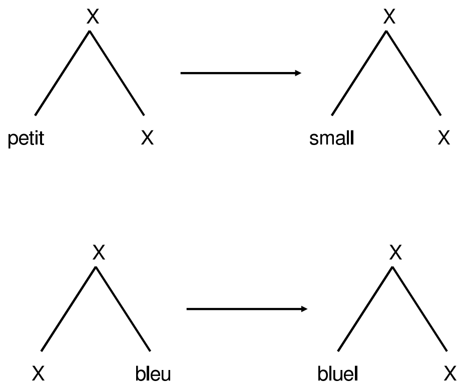


Figure 2.9: Example of hierarchical rules (French-English treelet pairs)

- better handling of ambiguities (opposed to deterministic rules)

This gives us an incentive to explore these directions, with the hope of overcoming the performance gap between rule-based and phrase-based systems on translation tasks such as defined for the Workshop on Machine Translation (Callison-Burch et al., 2007, 2008).

Indeed the low correlation with human judgement of the BLEU scores in the case of rule-based systems (Callison-Burch et al., 2006) might indicate a qualitative difference with the statistical outputs. Specifically, Ueffing et al. (2008) report system combination experiments which make use of various phrase-based systems along with a single hybrid system. In a combination system that aims at merging outputs from various systems, they experiment with dropping each system in turn. The outcome shows the largest loss in final translation quality when the one system with a rule-based component gets dropped.

2.4.1 Statistical Models with Linguistic Rules

While word- and phrase-based models originally are based on surface words only, syntactic models aim at representing the translation process at a more deeply structured level. A major motivation lies in the need to constrain the target language produced to be more grammatical, since this is a major weakness of phrase-based models. It usually means that either the source or target, or both, are represented by a tree structure (Figure 2.9).

A formal description of tree translation is possible using *synchronous context-free grammars*. Such grammars build two trees in parallel, one in the source language, the other in the target language. This kind of approach has received more and more attention since for example Wu (1997), who introduced syntax into the SMT framework, called Inversion Transduction Grammars (ITG), a category of bilingual context-free grammars that allow for reordering and probabilistic modelling. Yamada and Knight (2001) proposed a decoder along with their source-to-tree syntactical model.

Grammar may involve a single non-terminal symbol such as in Chiang (2005), or make use of linguistic syntax as in Galley et al. (2004). The question of how much the rules learnt may be human-understandable or not is of particular interest from the point of view of hybrid systems. This is evoked in Galley et al. (2004) who uses an English grammar not far from what rule-based systems use, as opposed to Wu (1997). Encoding translation rules within the framework of linguistic syntax offers the possibility to manually investigate the translation process, possibly, as described in DeNeefe et al. (2005) for the system of Galley et al. (2004).

More investigation is still needed to better understand the differences, and possibly the complementary relation between phrase-based and syntactical models, as done in DeNeefe et al. (2007).

Work on extracting (transfer) rules from a parallel corpus started from Moore (2001), Menezes and Richardson (2001) and Richardson et al. (2001). Quirk et al. (2005) made a breakthrough by learning dependency treelet translation pairs which showed to be able to produce better translation quality than state-of-the-art models (15 % better than MSR-MT and even 5 % better than Pharaoh), at the cost of speed. Quirk also introduced a decoding algorithm, while there was none in the original MSR-MT system.

Galley et al. (2004) proposes to extract syntactic transfer rules from parallel data, using a linguistic syntactic tree of the source language of the noisy channel model (that is to say, the target language when translating). This should be contrasted by Wu (1997) and Chiang (2005) which rely on a simplified, formal grammar. Instead, Galley motivates his work by the rule-based system framework and the analysis of the coverage of real human translations by syntactic models performed by Fox (2002).

Lavie (2008) describes a system that combines a morphological analyzer and synthesizer with synchronous context-free rules. These rules (both general and lexical rules) may be either manually created or acquired from parallel corpora. Ambiguity is managed by outputting a target lattice that reflects ambiguities (restricted to monotonic

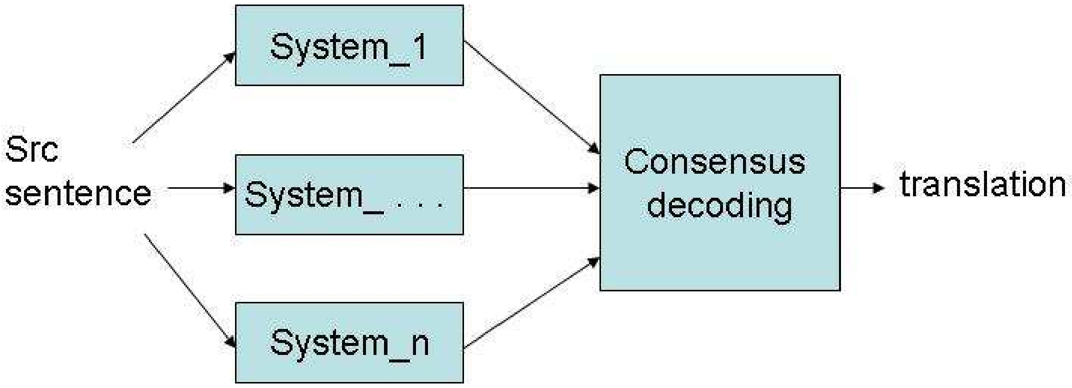


Figure 2.10: Consensus decoding

variations) which are decoded by a beam-search decoder in a second phase. Results on the standard Europarl data set are reported in the WMT 2008 competition and are slightly higher than the general-purpose (untrained) rule-based systems but much lower than the baseline phrase-based system.

Koehn and Hoang (2007) propose an extension to the phrase-based model allowing for the use of annotated tokens and the insertion of intermediary levels in the translation workflow, so as to deal with the issue of translating morphology. This introduces a *generation* step not unlike the synthesis step of rule-based systems.

However, rules entered in an RBMT system remain of a different nature. First, they use rich features that include linguistic syntax, morphology and sometimes semantic tagging. Second, they are manually entered and, for the most part lexicalized, would fail to be learnt reliably by a statistical learner, for lack of data.

2.4.2 System combination

Multi-engine systems form a first category of shallow experimental set-ups of hybrid systems. A multi engine system consists of a set of different MT systems, plus a combination system (Figure 2.10). Given some input, the combined system must be able to compose the best possible output by either selecting or combining the outputs of the different MT systems. An improvement is expected over the overall “best” individual system as long as these systems are not too closely correlated. This method does not require any modifications of the systems involved nor any attention to be paid to their interface. What is at stake is how to combine them in order to optimise the final output it produces. In the perspective of more integrated combinations of the different techniques and resources involved in these systems, a multi-engine experiment may give us insight about the value of each system.

Hogan and Frederking (1998) describes one of the first experiments which proves such a system may actually bring an improved translation performance. This setup however only picks up the best output out of those proposed by the different MT systems according to a target language model. In Bangalore et al. (2001), the outputs are combined together at the level of word sequence, forming a lattice from which the optimal path is found by considering, for each segment for which a choice is to be made, both the votes of all systems and the cost given by a language model. Combining of the multiple outputs at different levels is explored in Rosti et al. (2007). One may note however some weakness in the lattice approach, from having to align outputs with differing word orders, sometimes at a wide distance that makes it difficult to be solved by local word-based models. Heuristics are proposed to solve this, using for example the most consensual hypothesis (according to an edit distance metric) as a skeleton around which to start building the lattice/confusion network. A common intuition for this type of setup to be effective is the low correlation of the involved systems to be combined. Macherey and Och (2007) explore how to produce a diversity of translation systems from a common baseline and the impact of translation quality of each system on the combination. Some other combination methods are explored in that paper, such as the use of a sentence level BLEU correlation matrix.

In the framework of this thesis, we would like to mention issues raised by a setup combining both statistical and rule-based systems. First of all, the use of n-best lists may not always be possible (not mentioning a lattice or even a forest output), for it is not to be expected that a rule based systems will be able to output multiple hypotheses and score them. A combination using corresponding “phrase-tables” could also be problematic for the integration of a non phrase-based system.

2.4.3 Using rule-based output as a training corpus

Statistical systems claim to be automatically trained, without any manual intervention other than designing and implementing the algorithms involved. However, since they require a parallel corpus, they do make use of a manual input, albeit a side-product of an independent social activity. As translation quality grows with the available amount of such data, some have tried to produce *artifical data*, using existing machine translation systems, either rule-based (Hu et al., 2007; Dugast et al., 2008) or statistical (AbduI-Rauf and Schwenk, 2009).

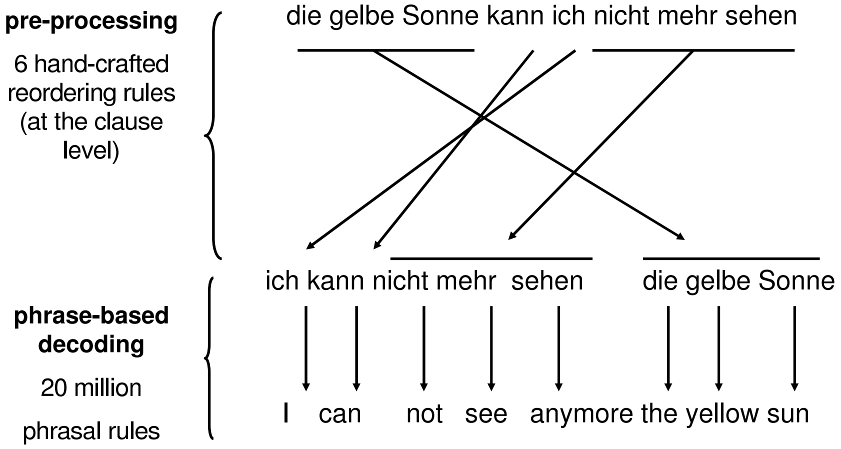

pre-processing
6 hand-crafted reordering rules (at the clause level)
die gelbe Sonne kann ich nicht mehr sehen
ich kann nicht mehr sehen die gelbe Sonne
phrase-based decoding
20 million phrasal rules
I can not see anymore the yellow sun

Figure 2.11: Clause restructuring following Collins et al. (2005)

Both attempts have shown to be successful. Wu and Wang (2009) compares three different pivot strategies in order to bridge a data gap between two languages using a third one. They make use of both SMT and RBMT systems, and show the RBMT systems bring a significant advantage on using solely the SMT system learnt from the available data.

2.4.4 Pre- or post-processing

Evaluated on specific domains with automatic metrics, phrase-based statistical models have proven to be ahead of rule-based systems, mainly because of their numerous unstructured rules (or ”phrase-pairs”). This kind of mapping of string sequences guided by string-sequence (”n-gram”) models of the target language is indeed in line with the automatic metrics (BLEU, Meteor, TER) based on string-sequences. Yet for some language pairs with remarkably different structures, such models seem to provide lower performances (for example, Chinese-English or German-English). Using rules as a pre-processing step to solve the long-distance reordering transformations from the source sentence to the target sentence is an option. The statistical model (phrase-based for instance) may then be left with dealing with local reordering only.

Methods such as of Collins et al. (2005) (Figure 2.11) propose to combine long-distance reordering by human-written rules before applying an SMT decoding. Alternatively, Xia and McCord (2004) use reordering rules that were automatically learnt

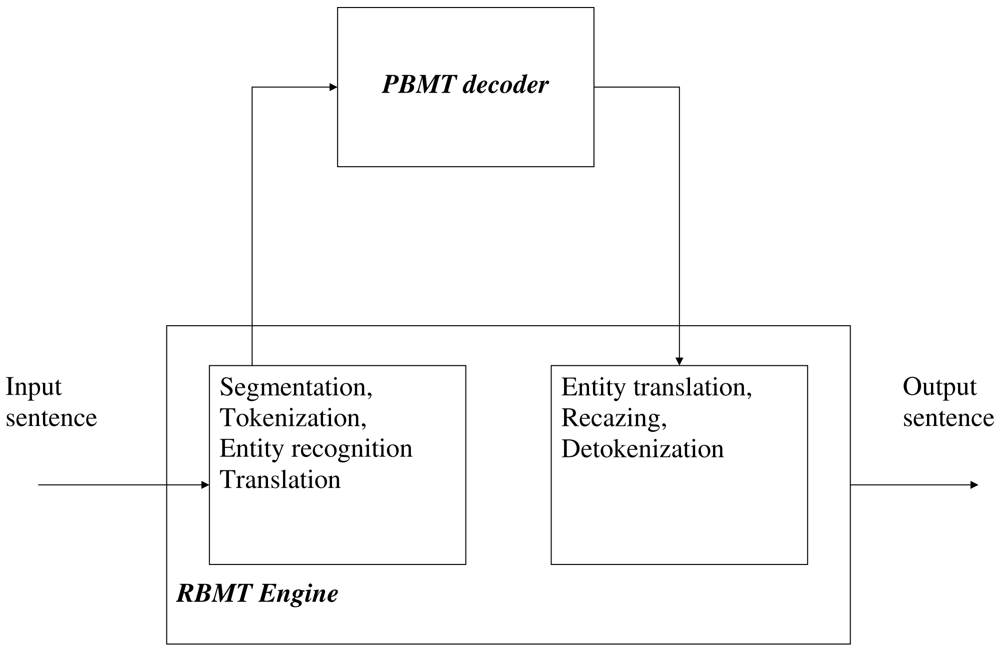

PBMT decoder
Input sentence
Segmentation, Tokenization, Entity recognition Translation
Entity translation, Recazing, Detokenization
Output sentence
RBMT Engine

Figure 2.12: The Statistical Post Editing workflow

from a corpus. In these approaches, rules are used to reorder the source sentence without translating it. In a second step, the reordered source sentence is translated by the SMT layer, which is supposed to perform only local reordering.

A more recent experiment envisaged such a serial combination as a post-editing task (Figure 2.12, where the SMT system was used to automatically correct the rule based output. The original idea of automating the post-editing task dates back to Knight and Chander (1994).

Knight and Chander (1994) argues for the construction of automated post-editors of a given language (”Anyone who has postedited a technical report or thesis written by a non-native speaker of English knows the potential of an automated postediting system.”). Post-editing of Machine Translation output is one of the applications of such a tool, all the more in the case of rule-based output, since errors will tend to be systematic, thus both tedious for human processing and easy to learn automatically. Good results are reported for article selection in English. Alternatively, also using a detached Post-Editing module, Llitjos and Carbonell (2006) propose to learn how to modify a rule-based system from the users’ post-editing of machine translated text. Developing the idea mentioned in Knight and Chander (1994), Simard et al. (2007a) built a statistical phrase-based post-editing model using a database of post-edited rule-based outputs from the Canadian job advertising bank. This idea was extended to using a parallel corpus (source and human translation) instead of truly post-edited rule-based machine translations. This appears both in Simard et al. (2007b) and in Dugast et al. (2007), the latter providing a qualitative analysis of the corrections performed by the post-editing layer. Ueffing et al. (2008) refined the combination of the rule-based and the statistical modules by using markup on the rule-based output.

These serial combinations were found out to be beneficial over both only-rules or only-SMT systems at the shared task on Machine Translation at the ACL2007 workshop, according to metrics both automatic and human. This was especially clear in the German to English task, for which it is known that long-distance dependencies are a major issue.

2.4.5 Terminology and rule extraction for a rule-based system

Managing lexicons, and especially *bilingual* lexicons is an especially hard task to handle manually, for reasons of size, domain and diachronic variations (Manning and Schütze (1999) : chapter 8 on Lexical Acquisition). Therefore, poor coverage is a major disadvantage of rule-based systems. Corpus-based methods may be used to ease the task of augmenting the dictionaries. It must be noticed that a rule based system generally makes use of structured, linguistically coded resources (whereas a ”phrase pair” in a statistical model may be formed of a pair of sequences of text which do not have to be linguistic phrases).

Extraction of terminology from a parallel corpus has been extensively studied. Kupiec (1993) is among the first to describe a pipeline for extracting dictionaries of noun compounds. Koehn (2003) gives a thorough investigation of the topic of noun phrase translation and extraction of noun phrase lexicons in particular. As far as extraction is concerned, the major variations between the different approaches lie first in the choice of either extracting first all monolingual terms to then find alignments between the monolingual terms or align chunks of texts (typically, extract a phrase table) which is then filtered to keep only the linguistically relevant ones. Daille et al. (1994) and Kumano and Hirakawa (1994) belong to the first category, when Itagaki et al. (2007) belongs to the second category.

A second kind of variation is the choice of confidence measures to evaluate the quality of a candidate entry. The frequency of the pair is the first obvious metric used to try sorting bilingual entries. Yet many more parameters, statistical or linguistic may be thrown in. Concerning the scale of those experiments, they do not reach the scale of current statistical phrase-based models, extracted for example from the Europarl corpus (Koehn, 2005).

Beyond terminology, we may want to extract sets of rules, for example disambiguation or transfer template rules (Och and Ney, 2004). It does not seem very common to extract rules from a corpus for the use in an RBMT system. In Carbonell et al. (2002), the paradigm is slightly different, since it is based on an ”elicitation corpus”, i.e. an on-purpose corpus, crafted in view of illustrating specific phenomena. Cicekli and Guvenir (1996) proposes some similar approach, based on learning by analogy.

On the other hand, works such as Galley et al. (2004) note the basic fact that many very general rules entered in rule-based systems are indeed relevant and that automatically learnt rules could gain from being linguistically motivated. On the aspect of readability of such rules, we may expect that the development method would at least differ deeply if the extracted rules are not readable. Can one use a method as described in DeNeefe et al. (2005), which proposes an interface for humans to review the automatically extracted rules, and possibly add some more manually.

2.4.6 Attempts at more intricate hybridizing techniques

In Dugast et al. (2008), the intuition that a simple N-gram model of the target language may bring most benefit to a rule-based output was confirmed. Indeed, the problem of local fluency for an MT system, well managed by current SMT systems thanks to their good coverage and the use of N-gram models (which actually model local fluency) can be seen as equivalent to the one encountered by a text generation system. This aspect is described in Langkilde and Knight (1998) in which a symbolic sentence generator outputs a word lattice from which the language model (bigram) ranks realisations.

Now, being able to combine a rule based system augmented with dictionaries, manually created or extracted from a corpus, with statistical decision modules as simple as Language Models to start with, we may begin to benefit from the complementary approaches. This is explored, though on a rather small scale, in Llitjos and Vogel (2007) in which an initial rule-based system with minimal dictionaries and rules is augmented with entries extracted from a parallel corpus. A language model is added to decide among different possible analysis of the source sentence.

From the point of view of a statistical framework, poor generalisation power is often identified as one of the major weaknesses of SMT systems. In order to gain generalisation power, a deeper level of representation of source and/or target language may be used. Linguistic factors may be used within a phrase-based model (Koehn and

Hoang, 2007), or syntactic models may be trained on parallel corpora. The latter kind may range from very lexicalised models (Yamada and Knight, 2001; Chiang, 2005) to models learning more general rules as (Galley et al., 2004). The latter chooses to use a linguistic, human-readable grammar of English to train a French-English or Chinese-English model. This kind of framework aims at getting closer to rules as more commonly understood in rule-based systems, which means allowing a certain level of generalisation for the automatically generated rules.

2.4.7 Research questions

This chapter has shown what the recent trends in research have explored so far. The noticeable persistence of rule-based systems motivates us to search for differences in their output. Can we then explain them through what we know of both inner workings of both approaches? Although recent developments have seen the introduction of formal grammars within statistical frameworks, such rules do not reach the complexity of common rule-based systems. Because they originate from a linguistic description of the translation process, advanced levels of knowledge are often used in rule-based systems. Syntactic analysis are combined with morphological features and semantic tagging of lexical items. Would this be possible to keep such advantages, while supplementing them with statistically extracted resources and statistical decision modules?

These questions will be examined in the next chapter.

Chapter 3

Comparison of Machine Translation Paradigms

The previous chapter intented to give an overview on the topic of combined techniques for Machine Translation (MT). In the following chapter, we will explore the details from the point of view of evaluation and qualitative analysis. What are the differences between translations produced by both approaches? We have given a stricter definitions of the terms “rule-based” and “statistical” and will now experiment with such systems. In addition, we will also build “black-box” combination systems. They are called “black-box” since they do not use any knowledge of systems’ inner workings. A qualitative analysis on our own experiments will provide us with a base for further work.

We intend to combine rule-based and statistical machine translation engines. We therefore need to evaluate strengths and weaknesses of both approaches. For that purpose, we sketch a qualitative description of the current approaches in research on MT. We then present experiments on a statistical model of the rule-based output and a statistical post-editing model that combines a statistical phrase-based layer with the rule-based output. Finally, we present experimental results on the comparison of such systems on a common dataset. An analysis of the errors in these experiments may guide the development of a hybrid model.

3.1 Evaluating Machine Translation output

In this section, we present various metrics (both automatic and manual) to assess the quality of MT output. We mention the reported bias of ngram-based metrics for ngram-based in-domain statistical models as opposed to general-purpose rule-based systems.

3.1.1 Manual evaluation

There are several issues in defining and using manual evaluations. First of all, we need to give annotators a common definition of what a good translation is. This definition may be vague and we choose to rely on relative judgements rather than absolute. We would even ideally want to evaluate separately the content transfer (*adequacy*) and the quality of the target language (*fluency*). A necessary condition for a human evaluation metric is to show good inter- and intra-annotator agreement. Such measures evaluate how consistent this definition may be. For instance, Callison-Burch et al. (2007) showed that *ranking* system outputs performed much more consistently than when annotators evaluate adequacy and fluency of system outputs. Finally, when using an evaluation metric, statistical significance tests are required to assess whether we may be able to draw conclusions on the relative quality of the systems from the obtained judgements.

3.1.2 Automatic metrics

Automatic evaluation metrics are an appealing substitute to manual evaluations, for those may be costly and time-consuming. They are relevant only if they have a good correlation with manual evaluations. This is shown by Papineni et al. (2002) for the BLEU metric, Banerjee and Lavie (2005) for METEOR, or Snover et al. (2006) for the TER metric. However, Callison-Burch et al. (2006) pointed out that the BLEU metric was biased towards statistical systems and did not allow for a fair comparison of rule-based and statistical systems (in the sense of correlation with the manual evaluation).

At least two factors are suspected causes of this bias: the use of n-gram counts for both the evaluation metric and the (phrase-based) model; and matching training and test material at the level of the word sequences which does not always imply a better translation, as judged by human experts. The former cause will overly reward correct word sequences without considering the grammatical structure and punishing very little the presence of noise (especially *missing* words as opposed to *spurious* words since they do not impact a precision-oriented metric such as BLEU) or important semantic mistakes (for example, negation). The second cause will overly reward specific jargon while give no credit at all to terms of equivalent meaning. Translation metrics are obviously not perfect. Nevertheless, they are used as objective functions to optimize statistical systems. As a consequence, any weakness of such a metric may be exploited by the optimization step and lead to a discrepancy between the metric score and the translation quality.

3.1.3 Error Analysis

For a qualitative understanding of the differences between system outputs, deeper error analysis is needed. Vilar et al. (2006) proposes a taxonomy of error types for manual annotation. Error analysis is however even more costly than manual evaluation of overall translation quality. Moreover there are issues of both intra (is the task clear enough for each annotator?) and inter annotator agreement (do they agree on the explanation of the observed errors?) There may well be more than one possible explanation for broken machine translation output. There have been also attempts at performing automatic error analysis (Popovic and Ney, 2007). Yet, they rely on imperfect tools such as part-of-speech taggers and automatic word-alignment and are far from giving the explanations we would hope for.

3.1.4 Related work: Qualitative analysis of MT technologies

Hierarchical (Chiang, 2005) and even syntactical (Zollmann and Venugopal, 2006) models have appeared in the field of statistical machine translation, bringing some diversity in a research otherwise dominated by phrase-based machine translation. This led to a few comparative analyses. DeNeefe et al. (2007) provides a comparative analytic view of syntax based and phrase based systems. Zollmann et al. (2008) compares the three approaches in terms of final translation performance, varying language pairs and data sizes. Birch et al. (2009) reports a qualitative comparison on the specific aspect of reordering. Auli et al. (2009) introduces the aspect of *induction error* in comparing a phrase-based and a hierarchical system. Making use of reference translations of the held-out test corpus, it itries to distinguish between translation errors due to the search algorithm and those inherent to the translation model.

We cannot help notice there is little or no comparative analysis including the rule-based systems. This is all the most surprising as such systems, though not customized on the domain, have been shown to compete favourably in a relatively open domain such as news (Callison-Burch et al., 2008). Thurmair (2004) presents a comparative study between a rule-based and a phrase-based statistical system. This study is however limited (for example the test set used to perform this comparison only contains 62 sentences), and the qualitative comments are not supported by any statistics on human judgements and/or corpus-based evidence.

3.2 Evaluated Hybrid Systems

In the following, we report on the qualitative and quantitative differences of rule-based, statistical and black box hybrid systems.

3.2.1 A phrase-based model of the rule-based output

Our first hybrid system is a phrase-based model trained on a synthetic corpus generated with a rule-based system. It aims at reproducing the rule-based output and does not use any target-language corpus.

A source language corpus is translated with the rule-based system, hereby creating a parallel corpus. A basic phrase-based system is trained from the parallel corpus and the target language model is learnt from this machine translation output.

See section 2.4.3 for the background of this approach.

The rule-based translation of the source text in the tuning set provides the parallel corpus to tune this system.

Such a model may be evaluated in two ways. For a given parallel test set, it should be compared with the rule-based translation of the source, since it was explicitly trained to reproduce it. It may also be scored against the reference human translation(s).

3.2.2 Using monolingual corpora in both languages

This model reuses the translation model learnt in the previously described setup. It however uses an n-gram language model trained on actual target language text, instead of machine translation output. The tuning stage is also performed using a source language corpus with a reference manual translation (the sole parallel corpus used in this setup). Beyond that, there is still no need for a parallel training corpus to learn the translation phrase pairs. It could also be trained on loosely comparable corpora. Therefore this still illustrates a case where massive parallel corpora are unavailable for training purpose.

3.2.3 A Post Editing Model of the rule-based output

We describe here a serial combination of the rule-based system with a phrase-based system that aims at correcting the rule-based output to match human, reference translation. See section 3.3 for more background of this approach.

The translation model is learnt to translate from the rule-based output to the reference manual translation. For this purpose, the source side of the parallel corpus is translated by the rule-based system, respecting sentence-level alignment. The language model training is not specific (identical to the model used in a standard phrase-based system). The parallel corpus for tuning is produced in the same way.

At runtime, the source sentence is first translated with the rule-based system, and then corrected by the phrase-based post-editing model.

Various improvements may be brought to this basic setup, taking advantage of the monolingual nature of the alignment, using subcomponents of the rule-based engines such as entity recognition and translation or informing the post-editing layer with confidence markup from the rule-based system. Experiments are described in Ueffing et al. (2008).

3.3 Experiment 1: statistical post-editing of the rule-based output

We describe here the statistical post-editing setup. Simard et al. (2007a) describe an experiment where they use manually post-edited machine translation outputs aligned with the original translation to train statistical phrase-based post-editing models with a standard beam-search decoder. Other experiments were then conducted, this time using direct human translations as reference instead of a true post-editing of the rule-based output by Simard et al. (2007b) and Dugast et al. (2007)). In Dugast et al. (2007) we provided a qualitative analysis of the edits performed by such a model.

3.3.1 Setup

PORTAGE (Sadat et al., 2005) is an implementation of the beam-search algorithm for phrase-based machine translation, very similar to the Moses system (Koehn et al., 2007). Statistical Post-Editing systems over the SYSTRAN rule-based output are trained following the setup described in Chapter 3.3, using either PORTAGE or Moses

	SYSTRAN PORTAGE	SYSTRAN Moses	SYSTRAN Moses
	En→Fr	De→En	Es→En
termchg_all	22%	46%	46%
termchg_nfw (unknown word)	0%	3%	1%
termchg_term	19%	42%	45%
termchg_loc	8%		
termchg_mean	-6%		
gram all	2%	4%	12%
gram_det	14%	2%	4%
gram_prep	2%	1%	5%
gram_pron	-1%	1%	4%
gram_tense	-4%	1%	0%
gram_number	0%	0%	0%
gram_gender	-4%	n/a	n/a
gram_other	-1%	None	None
puncdigitcase	1%	-1%	-1%
wordorder_short	-1%	1%	1%
wordorder_long	0%	None	1%
style	1%	3%	2%

Table 3.1: Relative improvements brought by the SPE system, ratio computed as such: (#improvements-#degradations)/#modifications

as the phrase-based decoder. Another minor difference between them lies in the fact that the system using Portage used both Europarl and the news corpus for training, while the other setup relied on Europarl data only.

Based on the data from these two experiments: SYSTRAN+PORTAGE (English↔French), and SYSTRAN+Moses (German→English and Spanish→English), we performed linguistic evaluations on the differences between the original SYSTRAN output and SYSTRAN+SPE output. The evaluation for English↔French was performed on the News Commentary test 2006 corpus, while the evaluations for German→English, and Spanish→English were performed on the Europarl test2007 corpus.

The following categories were introduced to qualify the changes:

- *termchg*: changes related to lexical changes

- *termchg_term*: terminology change not affecting Part-Of-Speech nor meaning
- *termchg_loc*: multiword expression or locution
- *termchg_mean*: terminology change altering the meaning
- *gram*: grammatical changes
- *gram_det*: change in determiner
- *gram_prep*: change in preposition
- *gram_pron*: change in pronoun
- *gram_tense*: change in tense
- *gram_number*: change in number
- *gram_gender*: change in gender
- *punct,digit,case*: change in punctuation, number entities or case
- *wordorder_local*: change in local word order
- *wordorder_long*: change in long-distance word order
- *style*: change in style

3.3.2 Results

Let us first take a look at the impact of the Statistical Post Editing on the SYSTRAN output. Table 3.2 displays the Word Change Rate (WCR: edit distance based on tokens, a Word Error Rate computed between the rule-based translation and its post-edition through a statistical model) and the ratio of sentences impacted by the statistical post-editing. On the one hand, it is interesting to note that the impact is quite high since almost all sentences were post-edited. On the other hand, the WCR of SYSTRAN+SPE is relatively small - this clearly shows the post-editing is not a complete reshuffling of the translation. The same insight is reinforced when reviewing a few examples (see Table 3.3) - the SYSTRAN+SPE output is ”reasonably” close to the original SYSTRAN

	Word Change Rate	**Sentence Change Rate**
SYSTRAN+PORTAGE English→Frencg (nc devtest 2006)	33%	98%
SYSTRAN+PORTAGE French→English (nc devtest 2006)	23%	95%
SYSTRAN+Moses German→English (nc test 2007)	35%	100%
SYSTRAN+Moses Spanish→English (nc test 2007)	31%	99%

Table 3.2: Impact on Rule-Based output and Automatic Evaluation of translation quality for the SPE experiments

output, and the SPE output sentence structure is completely based on the SYSTRAN output.

The qualitative analysis (Table 3.1) shows that the editing, however globally positive in terms of translation quality, results from improvements and degradations of diverse proportions according to the qualitative category.

Table 3.1 shows that the most improved category is the ”termchg” which corresponds to a local choice of word sense or alternative translation of words and locutions.

Among the grammatical categories, ”gram_det” and ”gram_prep” are improved but the improvement over degradation ratio for this last category is very low (it shows global improvements but there are many unacceptable degradations).

No ”long-distance” restructuring is observed and local reordering is negative for English to French and relatively negligible for other language pairs.

For the French target, morphology is a major issue. And indeed there is no mechanism in the SPE that would provide any control over morphology.

3.4 Experiment 2: comparing competing models on the same dataset

In this section, we describe an experiment where a variety of models are trained (on the same corpora when applicable, this is of course not the case for the rule-based system) and compared, using both automatic and human evaluation. An error analysis is also performed that highlights qualitative differences between the different setups.

Source	**SYSTRAN**	**SYSTRAN +SPE**
*Englis*h	*French*	*French*
Monetary policy can be used to stimulate an economy just as much as fiscal policy, if not more, in election years, which politicians will always want to do.	La politique monétaire peut être **employée** pour stimuler **une** économie **juste comme beaucoup que** la politique fiscale, **sinon** plus, **en** années d'élection, que les **politiciens** voudront toujours faire.	La politique monétaire peut être **utilisée** pour stimuler **l'**économie, **tout comme** la politique fiscale, **pour ne pas dire** plus, **dans les** années d'élection, que les **hommes politiques** voudront toujours faire.
German	*English*	*English*
Fortschritte der 12 Bewerberländer auf dem Weg zum Beitritt	Progress of the 12 **applicant** countries **on the way** to **the entry**	Progress of the 12 **candidate** countries **along the road** to **accession**
Spanish	*English*	*English*
En una perspectiva a más largo plazo, habrá una moneda única en todo el continente.	In a perspective **to more long term**, there will be a **unique** currency **in all the** continent.	In a **more long-term** perspective, there will be a **single** currency **for the whole** continent.

Table 3.3: Examples illustrating the effect of the statistical layer

Type	Number of rules
Single words	60 k
Phrase and disambiguation rules	100 k

Table 3.4: Number of hand written lexical rules in the SYSTRAN Rule-based System

3.4.1 Training data

We used version 3 of the Europarl corpus (Koehn, 2005) for the French-English language pair. Only the parallel data was used, whose size is 1.3M sentences with about 30M running words. Language models were trained using the target side of the training parallel corpus only. Tuning tests are identical except for the RELEARNT1, where the target side is replaced by the rule-based translation of the source-side. We score RELEARNT1 on both the rule-based output (against which it was both trained and tuned) and the actual manual reference translation (see further on Table 3.6).

3.4.2 Systems

We compare five different systems on this domain. We use the SYSTRAN engine as a rule-based system (*RBMT*). For the French-English language pair, we provide rough estimates of the number of manually coded lexical rules in table 3.4, though it is obviously difficult to compare this with the size of the corpus used to train the statistical models or the number of ”rules” (phrase pairs) that were learned from it. We trained a baseline phrase-based system using the training data in this domain (*PBMT*).

We then compare these systems against the three hybrid models described in Section 3.2. We trained systems using the rule-based output translation, either to reproduce the rule-based system (*RELEARNT1*) or enhanced with a language model trained on the reference manual translation of the source side (*RELEARNT2*). Finally, we train a post-editing model on this same data (*SPE*).

All evaluations are done at the sentence level.

3.4.3 Metrics

We choose to evaluate systems using the BLEU metric (Papineni et al., 2002), on the ”devtest” corpus of Europarl, as provided in the WMT competition. Results are given in table 3.6. The results show that, when considering the automatic scoring solely,

System	Rule-based, statistical or both	Phrase table target language	Language Model
RBMT	RULE	none	none
PBMT	STAT	EN	EN
RELEARNT1	STAT	SYS_A_EN	SYS_A_EN
RELEARNT2	STAT	SYS_A_EN	EN
SPE	BOTH	EN	EN

Table 3.5: Systems differences

the rule-based system performance can be reproduced in this domain, with this level of datasize.

3.4.4 Manual evaluation

As for manual evaluation, we choose to rely on the simultaneous ranking of the five systems, for this has been proven to provide good agreement among annotators, (Callison-Burch et al., 2007). This evaluation requires the judge to attribute a rank to each translation, starting from the one(s) judged best, with possible ties. 25 different annotators, all with a minimal knowledge of French performed 110 judgements in total. Results are given in table 3.7. The statistical model of the rule-based output (RELEARNT1) does indeed reach a score very close to the system it is aimed at reproducing. The Post-Editing model gets a slightly higher score than the phrase-based one, which might indicate that at this size of data, the rules contained in the RBMT system still help in getting a higher coverage of the source sentence. The intermediate system (RELEARNT2) gains a major part of the gap in BLEU points just by using a Natural English language model and tuning set.

Beyond overall quality we would like to identify specific mistakes made by each of the five systems we evaluate. We define a set of error categories as listed in Table 3.8. This is a slightly simplified breakdown of error types as suggested by Vilar et al. (2006). We are aware that such a task is difficult to define and provided annotators with the following guidelines:

- The span of each error is left to the (linguistic) judgement of the annotator. For example, ”safety meeting” in place of ”security council” should be judged as one error only, for ”security council” may be one lexical choice.

- Each erroneous sequence may qualify for more than one category. For example.

System	**T**uning final BLEU (%)	**T**est BLEU (%)
RBMT	No tuning	21.3
PBMT	30.0	29.9
RELEARNT1	84.9 †(20.5)	20.9
RELEARNT2	26.7	26.6
SPE	31.9	31.8

Table 3.6: Systems Evaluations: BLEU automatic metric.

†this tuning was done with RBMT translation as reference, score on real reference is given in brackets.

	RBMT	*PBMT*	*RELEARNT1*	*RELEARNT2*	*SPE*
RBMT	-	0.50	**0.64**†	**0.55**	0.27†
PBMT	0.50	-	**0.57**	**0.62**†	0.32†
RELEARNT1	0.36†	0.43	-	0.44	0.23†
RELEARNT2	0.45	0.38†	**0.56**	-	0.23†
SPE	**0.73**†	**0.68**†	**0.77**†	**0.77**†	-

Table 3.7: Systems Evaluations: manual evaluations. Ratio of sentences when system on row was judged to be (strictly) better than system on column. Ties were excluded from counts. †indicates results significant at the 5% level according to the Sign test.

Type of error (abbreviation)	Definition
MC	Missing Content
MO	Missing Other (i.e. grammatical word)
TCL	Translation Choice (content, lemma)
TCI	Translation Choice (content, inflection)
TCO	Translation Choice (other)
EWC	Extra Word Content
EWO	Extra Word Other (i.e. grammatical word)
UW	Unknown word
WOS	Word Order, short (distance < 3)
WOL	Word Order, long (distance ≥ 3 words)
PNC	Punctuation

Table 3.8: Translation Error Types

”chats” - ”mouse” is both a translation choice error (a correct translation would be ”cat”) and an inflection error (this should be plural, hence ”cats”).

- Content words are considered to be restricted to nouns, verbs, adjectives, adverbs. One may consider that some prepositions bear content too. We choose however not to classify them as content words. Other non-content words are determiners, conjunctions.

- As far as Translation Choice (Lemma) is concerned, favour meaning over style. Even if a translation choice could have been better style-wise, this should not be considered an error if the meaning is correct.

A simple online interface eased the recording of these judgements, taking on average three minutes to analyse errors of all five translations for a same source sentence. gives the total number of error for each category and system.

3.4.5 Results

Rankings induced by the BLEU metric and the manual evaluation all agree but for the pairs (RBMT vs RELEARNT2) and (RBMT vs PBMT). Comparison of tables 3.6 and 3.7 seems to indicate that having tuned RELEARNT2 with the in-domain n-gram Language Model tended to game the BLEU metric without a corresponding high jump

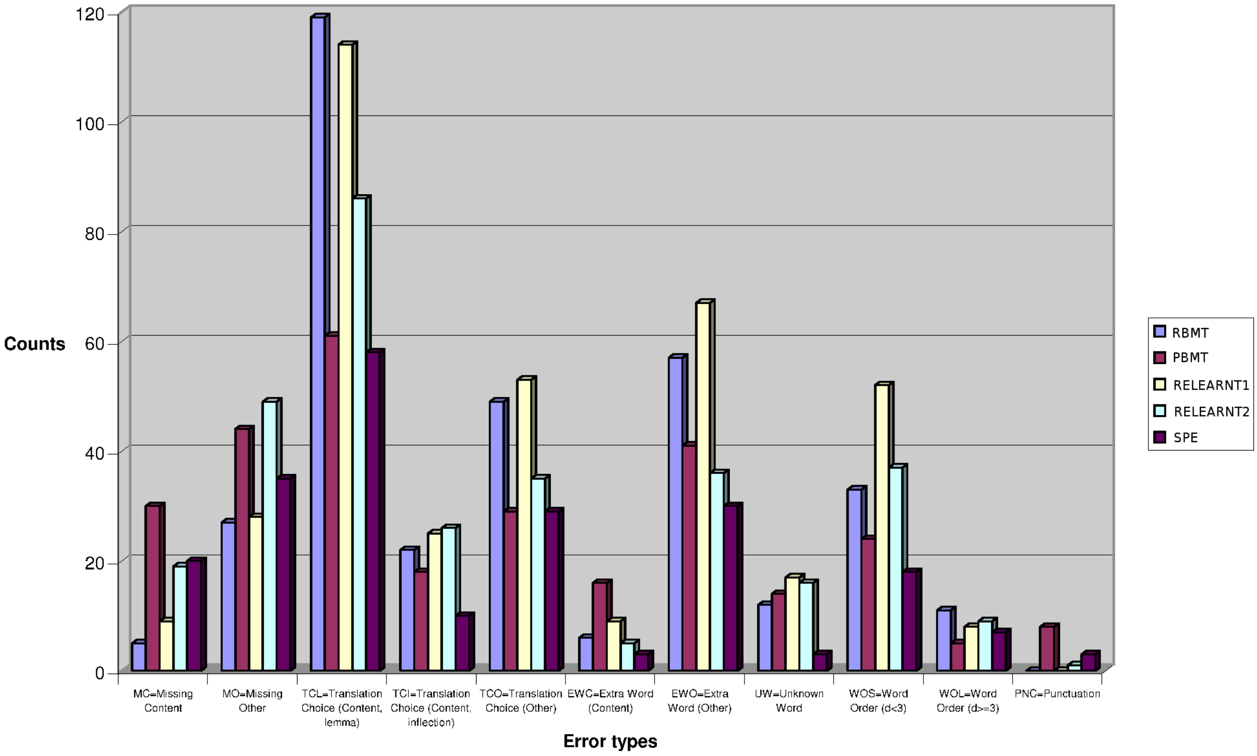


Figure 3.1: Error counts by category

in actual translation quality. Moreover, though getting only a slightly higher BLEU score, RBMT is still better ranked than RELEARNT1, its black-box re-engineered model. This may give an additional indication that a rule-based system is under-rated by the BLEU metric, regardless of its domain adaptation. In conformity with their largely superior BLEU score, PBMT and SPE systems, which were both trained and tuned on domain reference translation, rank at the top.

3.4.5.1 Mapping error types to system features

Let us take a look at the detailed breakdown of error categories in Figure 3.1 to identify how well the different systems perform on each error type.

In the categories of deleted words (*MC* and *MO*), the rule-based system has the smallest amount of errors. In contrast, the combination of the language model and discriminative tuning on the BLEU metric using a human reference translation seems to hurt (37 to 68 errors for RELEARNT2 compared to RELEARNT1). The use of a non-artificial phrase table adds up more errors for content words (PBMT). From this we may get the insight that noise (*superfluous* and *deleted* words) are first due to the tuning of a language-model-driven decoding to the BLEU metric, which is precision rather than recall oriented. This is confirmed by the greater amount of deletions as compared with extra words for systems PBMT, RELEARNT2 and SPE. Another cause for this problem seems to be the word alignment of noisy natural language parallel text. Systems using artificial parallel data (RELEARNT1 and RELEARNT2) or monolingual parallel data (SPE) display fewer errors of this sort.

We can see that the overgeneralization of the rules in the rule-based systems leads to many unnecessary grammatical words. The use of a language model trained on natural English text of the domain reduces this problem (from 67 to 36 errors between systems RELEARNT1 and RELEARNT2). The choice of lexical translation (*TCL* category) is clearly the main weakness of the rule-based engine. As shown by comparing systems RELEARNT1 and RELEARNT2 in this respect, the language model allows to bridge about half of the gap with the phrase-based model in this category. The other half (comparing RELEARNT2 and PBMT) is bridged by the use of natural phrases.

Concerning inflection, we notice that the post-editing model gains a significant reduction of errors compared with otherwise similar performances of the other systems. The choice of grammatical words, another main weakness of the rule-based model is also mostly improved by the language model (systems C to D), and less importantly by the phrase table.

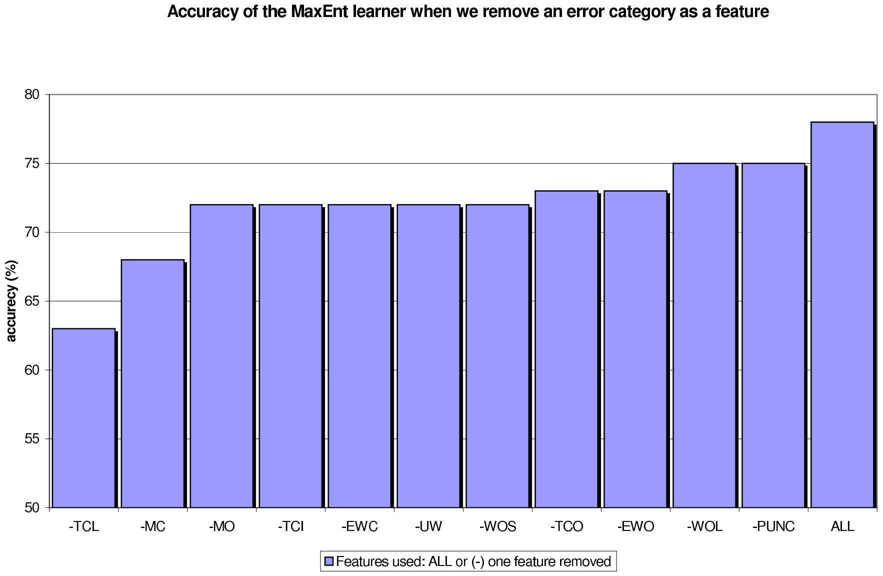


Figure 3.2: Accuracy of the MaxEnt learner when dropping one of the error category as a feature

As far as unknown words are concerned, all systems seem to be equivalent for this domain, except for the hybrid post-editing which seems to gain coverage (from 12 to 3 errors compared with the rule-based) from the union of the existing dictionaries in the rule base and the post-editing phrase-table extracted from in-domain text.

Short-distance word order is better handled by the phrase-based model than the rule-based (24 and 33 errors respectively), with the language model bringing most of the improvement. Even at a wider span (distance of at least 3 words), the phrase-based system performs better, possibly thanks to longer phrases in the phrase-table, as the difference with RELEARNT2 might show. Unsurprisingly, punctuation is an issue for the statistical models only.

Finally, the post-editing model seems to qualitatively equal or gain on both the rule-based and the statistical engines, except for the deleted words categories. We would now like to explore how these types of errors relate to the overall perceived quality of the translation.

3.4.5.2 An attempt at mapping error types to relative quality

We would like to evaluate the relative informativeness of each of these types of error in the judgement of the relative translation quality. The basic assumption is naturally that a translation with more errors is likely to be judged worse globally. Except that some errors may matter more than others.

We dump the data we collected in both the human evaluation and the error analysis as training data for a classification problem. Samples are pairs of translation outputs, whose features are the differences in the number of errors of each type and the label is chosen in the set $\{$BETTER,WORSE$\}$.

Figure 3.2 presents the accuracy of the Maximum Entropy classifiers we train. We run an initial training with all the error categories, then run more trainings with a feature dropped each time. We hope this way to spot how much each type of error matters to predict which translation is better.

The lexical translation errors are both the most frequent type of error (Table 3.1)and the feature that hurts most prediction when dropped. However, the *Missing Content* type of error comes second in this respect in spite of much fewer counts.

As a conclusion, while the most impacting categories seem to be those better handled by the phrase-based system (translation choice and short-range word order), the noise introduced by the statistical models (deleted and extra words), though a less frequent type of error, has a determining impact on the global judgement on the translation quality. On the other hand, very frequent errors such as the choice of grammatical words (or the generation of purious grammatical words), if likely to have an important impact when using string-based metrics such as BLEU, seem to matter much less.

3.5 Discussion

In the present chapter, we described and studied characteristics of different approaches in Machine Translation. We introduced system types that are hybridisations of the two paradigms we focus on: rule-based and statistical machine translation. A qualitative analysis of a straight-forward successful hybrid setup gave us a first insight of which strengths of each approach could ideally be combined. On a given corpus and for one language pair we presented a comparative study of rule-based, statistical phrase-based and these initial hybrid systems that led us to more specific conclusions on what differentiates the two approaches and how this relates to translation quality.

First and foremost, we reproduced anterior results which show how automatic evaluation algorithms are biased towards phrase-based statistical machine translation versus rule-based machine translation. In our study, the quality of both approaches appeared to be fairly comparable when evaluated manually.

Secondly, using a simple grid of translation errors, we could see that the major advantage of the phrase-based system in comparison with the rule-based lies in the learning and context-based choice of phrasal lexical rules. On the other hand, the strength of the rule-based system lies in the realatively stable structure of the output sentences, while avoiding the noise of missing or spurious items that is the trademark of a statistical system.

All this will guide our further development of hybrid systems and is the ground for our following chapter on the extraction of phrasal lexical rules.

Chapter 4

Automatic Lexical Rule Acquisition for Rule-Based Systems

The previous chapter intended to ground our crafting of a combined system on qualitative observations. Our comparative study indeed pointed to us some linguistically characterised aspects of translation where most of the improvement is expected from the introduction of statistical algorithms and models.

Now going beyond the black-box hybrid systems, we aim at porting the strongest identified feature of the statistical approach to rule-based systems: the automatic acquisition of translation rules. More precisely, we focus on the acquisition of a particular type of rules: lexical phrasal rules. In a rule-based system, they constitute both the least stable and the most time-consuming component. We therefore present a method for dictionary extraction for a rule-based system. Further on, we present a pruning algorithm that aims at optimizing translation quality while using the potentially noisy extracted set of rules.

4.1 Dictionary extraction

4.1.1 Motivation

We deal here with one of the expected benefits of using statistical techniques in rule-based systems: automating what had otherwise been a manual task. Entering more words or terms (syntactic phrases) in the dictionaries has always been the most costly and tedious task in developing the rule-based translation engine, either to improve coverage or to adapt to a specific domain. Moreover, Chapter 3 showed that lexical

	English	**French**	**Benefit**
1	big park	grand parc	local context
2	private bank	banque privée	local context
3	left bank	rive gauche	local context
4	fig leaf	feuille de vigne	non-literal phrase
5	fraud scandal	scandale en matière de fraude	non-literal phrase
6	freight traffic	traffic de marchandises	provides syntactic disambiguation
7	to let off steam	décompresser	non-literal phrase

Table 4.1: Examples of phrasal entries

coverage (and disambiguation) was the most important qualitative advantage of statistical systems. Other experiments on hybrid systems such as those by Eisele et al. (2008) confirm this. The use of phrasal entries was also shown to be an important move when introduced in statistical systems, as compared with word-based models. Phrasal entries capture local context that provides immediate disambiguation, capture non-literal translations and finally may simplify the syntactic analysis of the source sentence. Examples in Table 4.1 illustrate these three aspects. Moreover, note that although there is evidence of a linguistic phrasal lexicon (Bannard, 2006), this type of resource (as a manually created dictionary) is not as easily available as simple-word dictionaries for a given domain.

4.1.2 Semi-automated addition of entries

As it is often the case, the rule-based system we use comes with a dictionary coding tool (Senellart et al., 2003) that allows the manual task of coding entries to be partially automated. It uses monolingual dictionaries (Table 4.2), morphological guess rules and weighted context-free local grammars (Table 4.3). The syntactic coding of lexical rules is necessary for the syntactic disambiguation phase in the system. Both side of the candidate phrase pair is parsed with a phrase grammar of the language involved. Whenever one side fails to be parsed or the syntactic categories of both sides do not match, the candidate entry is rejected. This feature allows a filtering of the automatically extracted phrase pairs. Moreover, morphological coding allows to generate target inflected forms according to both inflection translation rules form one language to the other and inflection tables which describe the morphology of the target language.

For example, the second rule illustrated in the table 4.3: $N_{+ZZC} \rightarrow < A >^0 < N :$

lemma	**part of speech**	**semantic tags**	**inflection code**
baptismal	A	$+EVENT+QUAL+RELA+RELIG$	A15
absorbance	N	$+ABS+MS$	N1
abound	V	$+AN+PREPR=(WITH,IN)+UINT$	V4
abroad	ADV	$+ADVVB+AN+PL+RADVA+REMOTE$	ADV

Table 4.2: Sample of the monolingual dictionary for English

rule	**headword index**	**arbitrary weight**
$N_{+ZZC} \rightarrow < N >^0 < N : *1_{-ZZC} >^1$	1	0.9
$N_{+ZZC} \rightarrow < A >^0 < N : *1 >^1$	1	1
$A_{+ZZC} \rightarrow < ADV >^0 < A >^1$	1	0.9
$V \rightarrow < V : *1 >^0 < CONJ >^1 < V : *1_{-REALW} >^2$	0	1
$ADV \rightarrow < ADV >^0 < CONJ >^1 < ADV >^2$	(none)	0.8

Table 4.3: Sample of the monolingual grammar describing English phrases. Conventions: N= noun; A=adjective; ADV=adverb; V=verb; CONJ=conjunction; +zzc=constituent; +realw=inflected form

$*1 >^1$ simply describes how an English noun phrase may be composed of an *adjective+noun* sequence. On the left part of the rule, the $_{+ZZC}$ index indicates this will be a Noun *Phrase*. Both words (Adjective and Noun) are identified by indices. They are used in the next column to specify the headword for the phrase. Moreover, the **1* symbol indicates that this part of the phrase should get inflected, according to the morphological features of the compound. In this case then, the *Noun* word would get inflected, while the *Adjective* word would remain constant.

The default coding rule has a phrase that inherits inflection and semantic features from the left-hand-side. The coding tool also allows the user to fine-tune the linguistic coding of an entry by correcting the automatic coding and/or enrich it with more features.

4.1.3 Extraction of a syntactic phrase table

We intend to extract linguistically coded phrase rules from a parallel corpus.

For that purpose, the extraction setup as depicted in Figure 4.1 starts from the parallel corpus from which a phrase table is built by the state-of-the-art procedure of using word alignments and heuristical phrase extraction. At this stage the ”phrases” are

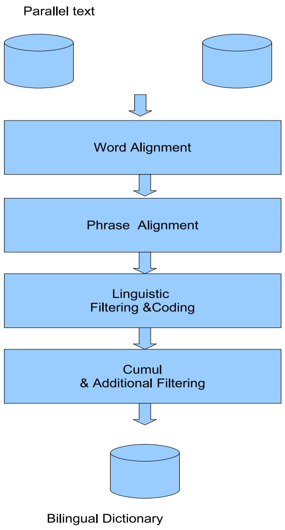


Figure 4.1: Extraction pipeline: from parallel texts to bilingual dictionary

plain word sequences, not necessarily linguistically motivated. They may also include part of speech annotations derived from the baseline RBMT system.

Each phrase pair is then processed by the dictionary coding engine, resulting in a pair of treelets such as illustrated by Figure 4.2. They have been produced by the weighted grammar we illustrated in Table 4.3. This is a limitation of the rule-based system to restrict the extraction to entries for which the target syntactic category is identical with the source category. Note that Koehn (2003) found for German-English that 98% of noun phrases could be translated as noun phrases. The extraction we perform is however not limited to noun phrases but also include verb, adjective and adverb phrases.

Some statistical features are attached to each phrase pair: frequency of the pair and lexical weights (Koehn et al., 2003) in both directions (see Table 4.4). As a bilingual entry may have various inflectional forms in the corpus, we sum over the lemma counts, from which we compute frequencies to perform filtering. We retain only the most frequent translation for each source phrase. If there are multiple highest frequency translations, we use the lexical weights as a tie-breaker.

4.2 Optimizing translation quality

The dictionary extraction method that we described does not consider the other types of rules in the rule-based system. Consequently, there is no guarantee, however accurate the method may be (as evaluated by the manual judgement of its entries) that the overall translation quality will actually improve. Note that most statistical translation

Cat	French	English	Freq	P(f/e)	P(e/f)
N	ancien *régime*	old *regime*	10	1	1
Adj	sans rapport	*unrelated*	3	1	0.5
Adv	dans l'absolu	*ideally*	2	1	0.4
V	*montrer* clairement	*make* clear	2	1	1
N	*accord* de base	rough *agreement*	4	0.67	1
N	accord de paix	peace agreement	31	0.32	1
N	accord de paix	agreement	23	0.24	0.11
N	accord de paix	settlement	17	0.18	0.21
N	accord de paix	peace settlement	14	0.15	1
N	accord de paix	peace deal	6	0.063	1
N	accord de paix	peace accord	5	0.052	1
Adj	à faible coût	low-cost	2	1	0.5
Adv	de façon explicite	explicitly	2	1	1
V	tirer parti de	take advantage of	2	1	0.09
N	grande devise de réserve	major reserve currency	2	1	0.5
Adj	en voie de développement	developing	10	1	0.625
Adv	dans un avenir proche	in the near future	5	0.63	0.71
V	taire les mauvaises nouvelles	conceal bad news	2	1	1

Table 4.4: Sample of extracted entries for French to English

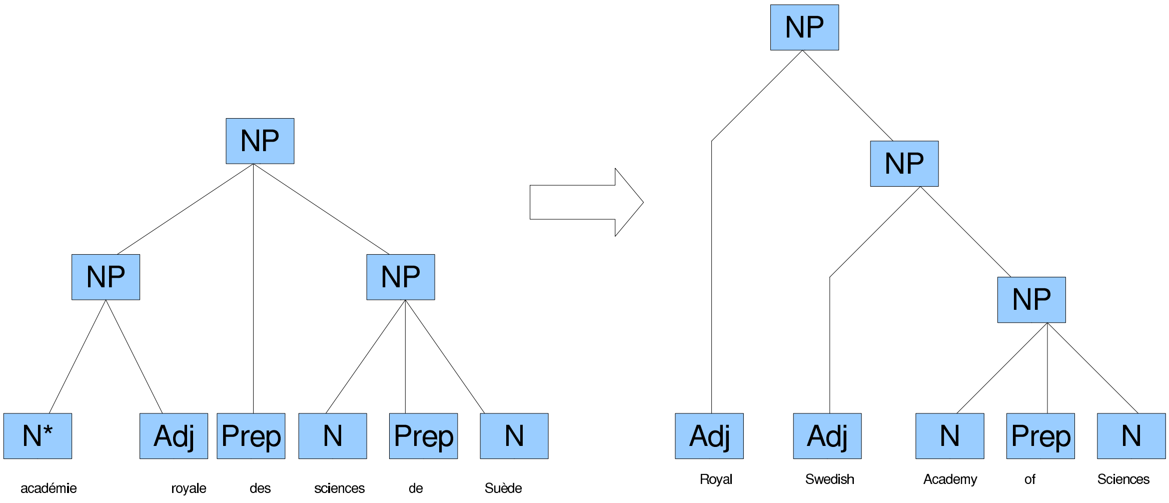


Figure 4.2: Example of extracted treelet pair

models, while extracting their rules with an independent procedure, resort to a discriminative tuning of global weights to optimize the final objective of translation quality. In the case of an augmented rule-based system, we have to deal with the connected constraints of fixed, unscored rules and the absence of a choice at runtime of which rules to apply. The problem we want to solve is thus to find the optimal pruning of the extracted lexical phrasal rules.

Indeed, the coding procedure, when applied to phrase pairs extracted from the corpus instead of manually entered entries, may generate rules that hurt translation quality. For instance, since the original rule-based system does not provide any means of exploring parsing ambiguities (a unique source analysis is produced by the rule-based parser), newly added (contiguous) phrasal rules may disable original rules and/or hurt the dependency analysis. It may also overrule existing, more sophisticated rules that had been manually entered. And finally, rules may just overfit the training data and fail to generalize.

We may define the optimal subset of the extracted dictionary with respect to a translation metric such as BLEU.

Algorithm 1 Dictionary Validation Algorithm

```
for n=1 to NgramMax do
    map all entries of some size n to parallel sentences in the training corpus
    translate training corpus with current dictionary
    for each entry do
        translate all relevant sentences with current dictionary, plus this entry
        compute BLEU scores with and without the entry
    end for
    Select entries with better/worse sentences ratio above threshold
    add these entries to current dictionary
end for
```

As an approximate (suboptimal) response to this problem, we test each extracted entry individually, starting from the lower n-grams to the longer (source) chunks, following Algorithm 1. For each sentence pair where the entry (of source span N) fires, the translation score (sentence level BLEU) when adding this rule is compared with the baseline translation. Rules showing only a single improved sentence translation or a ratio of improved against regressed translations below a given threshold (arbitrarily set at 1.3) are pruned out. The remaining entries are added to the system; providing a new baseline for the next iteration where rules of source span N+1 will be tested. The BLEU score of a held-out development set is computed at each iteration of adding longer-spanning rules.

The validation algorithm is along the lines of Imamura et al. (2003). However, our setup does not implement either cross-validation nor any greedy attempt at finding a subset of rules closer to the optimum. We will instead, later on, perform an oracle experiment to compare the upper bounds in that direction with the upper bound of a decoding or disambiguating module combined with rule extraction.

4.2.1 Evaluation and error analysis

The goal of this work is to improve the translation quality of a rule-based system by adding a large (at least, comparable with the scale of the existing manually created lexical resources) dictionary of word and phrasal entries. We want to check first of all the quality of the dictionary itself. We then want to evaluate and qualify the effect of this dictionary when used within the translation engine.

4.2.1.1 Evaluation of dictionary extraction

Let us first look at the intrisic quality of the dictionary. The criteria for the correctness of a dictionary entry are as follows: (1) both word sequences must be phrases of the grammatical category it has been assigned, (2) the lemma and inflectional codes have to be correct, (3) the headword must be correctly identified on both sides (while we do not require the local parses to be fully correct, see Figure 4.2 where the French parse is wrong) and finally (4) the target phrase must be a plausible translation of the source phrase.

We want to measure not only precision (the rate of good entries among the extracted set) but also recall (the rate of correctly extracted entries among the extractable correct entries in the data). Recall especially matters since such a setup for automatic extraction of entries is motivated by its ability to leverage lexical coverage.

In order to avoid repetitive human evaluation for the various experiments we may run, we create an automatic metric for this purpose. A subset of 50 sentence pairs from the training corpus is randomly selected. This constitutes the Gold Standard training set. From this subset, human annotators are asked to extract and code all the relevant bilingual phrasal entries. Preexisting tools for translation memory review and dictionary coding are used for this purpose. This constitutes the Gold Standard dictionary.

We make sure during the training process, that the original sentences can be traced back from the extracted entries. This allows Precision, recall and consequently F-Measure to be computed by comparing this extracted subset with the Gold Standard dictionary. This assumes that all entries in the Gold Standard are good entries and all good entries that we can possibly extract are contained in this Gold Standard dictionary.

In addition to this evaluation in terms of precision and recall, we also perform manual error analysis on a random sample of a hundred dictionary entries. The main categories of errors are:

- alignment error: one or both sides of the entry have been truncated

- syntactic category error: a verb phrase has been wrongly parsed as a noun phrase, for example

- coding error: lemmatisation or identification of the headword is wrong

4.2.1.2 Evaluation of translation

Given a certain quality of a dictionary, we now face the question of how much the dictionary improves translation quality. We evaluate translation quality with an automatic metric (Papineni et al., 2002) and human judgement. Although the BLEU metric has been shown to be unreliable (Callison-Burch et al., 2006) for comparing systems with such different architectures as rule-based and statistical systems, this does not discard its use for comparing two versions of a given system.

As far as human judgement is concerned, in accordance with the findings of recent evaluation campaigns (Callison-Burch et al., 2007), we choose to rely on a ranking of the overall quality of competing outputs. In addition to evaluation, we also perform a human error analysis on a random sample of a hundred sentences. This task consists of comparing the translation output when adding all the extracted rules with the baseline translation and trying to identify reasons for possible deteriorations or improvements.

4.3 Experimental results

4.3.1 Dictionary extraction

Our basic dictionary extraction configuration follows the pipeline described above. All phrases up to a length of 6 tokens are kept. Source phrases of different parts of speech are treated separately. Only the most frequent translation for each source phrase is kept. In the case of ties, the best aligned translation according to the IBM1 word based model score is chosen.

The Europarl parallel corpora for English-French is used for training and validation. The progress of translation quality as rules are added is monitored on the held-out *devtest2006* corpus, while final evaluation is done on the *test2008* test set.

Precision, recall and F1 measure obtained for dictionary extraction are displayed in Table 4.5. The baseline F1 score is relatively low. Using the rule-based Part-Of-Speech tags to enforce the linguistic coding of the phrase ensures a higher precision and F1 measure. Such tagging is provided by the syntactic analysis from the RBMT system. Without it, all possible taggings of each token are considered with different weights independently from context, resulting in different (weighted) parses of the phrase, out of which the best scored one is picked.

We are aware that precision may be underestimated, because the human annotator may have forgotten entries. And recall may be overestimated for the same reason. We however use it to compare setups: here, without or with the use of part-of-speech tagging (obtained from the baseline translation engine). We also evaluated the precision of the extracted entries for each syntactic category by manual judgement on a random sample (Table 4.6). The 64% precision for noun phrases when using the part-of-speech tags is similar to the result obtained by Itagaki et al. (2007) before filtering.

Retaining only one translation per source phrase for a given category, we extracted approximately one million entries in both setups. The two most important sources of extraction error are word alignment (35%) and category (45%). The remaining 20% come from coding errors (wrong headword or lemma). The first one comes from GIZA misalignments which may lead to a truncated source or target sequence. The ”Category” error type identifies with parsing errors of the local grammars used to code each monolingual phrase (Figure 4.2) that lead to an incorrect phrase category. Entry #2 of Table 4.7 for example should in reality be ”development *in connection with* the Millenium Goals”-”développement *dans le cadre des* objectifs du Millénaire”. The other two remaining types of errors involve linguistic coding. The identification of

Setup	**Precision**	**Recall**	**F1**
baseline	32%	65%	41%
+ enforced Part Of Speech from rule-based tagging	46%	49%	45%
+validation	52%	n.a.	n.a.
+ p.o.s.+validation	71%	n.a.	n.a.

Table 4.5: Automatic Evaluation of dictionary extraction w.r.t. the Gold Standard

% correct phrases in category	**baseline**	**+ p.o.s.**
noun	56	64
verb	52	64
adjective	38	38
adverb	36	38

Table 4.6: Human Evaluation of dictionary extraction (most frequent meaning only)

headword is crucial because, as a default rule of the coding engine, the entry will inherit its properties from it, especially determining its inflection pattern. This consequently matters for both the sake of coverage of inflected source phrases and generating the correct inflected target phrase.

4.3.2 Application of extracted dictionaries

The baseline system is the original rule-based system. We compare it with the augmented system that uses additional validated rules extracted from the corpus.

Table 4.8 shows the most frequent causes of deterioration when adding all the rules. Only a part of the causes for deteriorations is due to extracted dictionary entries that would be manually judged incorrect. The other reasons of decreasing translation quality have to do with either part-of-speech ambiguity, negative interaction with the dependency analysis, and the lack of a mechanism for translation choice or interaction

#	**Err. type**	**English**	**French**
1	Alignment	*correction* in the stock	*correction* des bourses
2	Category	*development* in connection	*développement* dans le cadre
3	Headword	controlling migration *flow*	*contrôle* des flux migratoires
4	Lemma	hand of the national authorities	main des autorités nationale

Table 4.7: Examples of extraction errors (headword is emphasised)

Type of error	errors
Syntactic Ambiguity (category)	19%
Syntactic Ambiguity (other)	21%
Wrong Translation (bad dictionary entry)	16%
Wrong Translation (inappropriate translation in context)	9%
Interaction With Other Rules	28%

Table 4.8: Translation deterioration Analysis on System2, original rule based system with all extracted rules

System	% BLEU	improved	worsened	equal
B	24.2	n.a.	n.a.	n.a.
S2	21.4	20%	69%	12%
S3	27.1	64%	22%	14%

Table 4.9: Automatic evaluation of translation quality and human evaluation of deterioration. NIST bleu on the test2008 dataset (realcased, untokenised output). B=baseline; S2=baseline+all rules; S3=baseline+validated rules

with the existing set of rules.

Figure 4.3 shows that the metric-based filtering of entries manages to improve the overall translation quality. It appears that the use of part of speech tagging did not improve the final BLEU score. This might be due to the combination of the lower recall (for a higher precision though) and the ability of the validation process to get rid of a higher number of bad entries in the other extracted set. Only 67k entries are finally retained at the end of this process. This compares to the pre-existing dictionary of around 300k entries, made up half of simple words and half of phrase entries.

Table 4.9 presents both BLEU scores and human evaluation of improvement or deterioration as compared with the baseline, non augmented system, for both augmented systems. When translating the 2000 sentences test set with the setup using the pruned set of entries, 3519 extracted entries were used (3486 unique), covering 12% of the source tokens. Table 4.11 illustrates discarded and retained entries while Table 4.10 shows two samples of compared translations.

We can see that, by adding more and more context-specific rules, we manage to rise the BLEU translation score on the held-out corpus by 3 points. This is contrasted by the behaviour of a non-filtered set of rules.

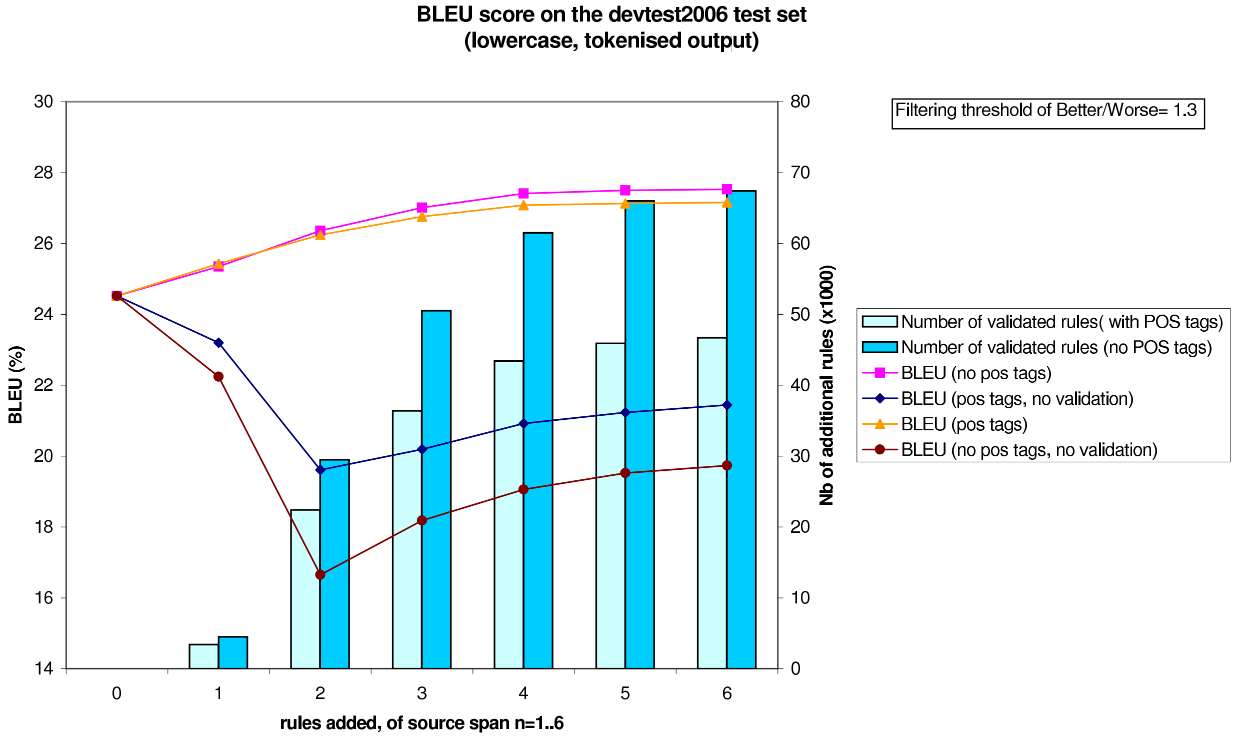


Figure 4.3: Progress of BLEU score at each iteration of the validation process

Source	Allow me also to say *at this point* that I have a great deal of respect for the citizens *of the central and eastern European countries* who, ten years ago, had the courage to go into the streets and start this process.
Reference	Mais qu'il me soit également permis de dire mon respect pour les citoyens des pays d'Europe centrale et orientale qui eurent le courage, il y a dix ans, de descendre dans la rue et qui ont contribué à mettre en branle ce processus.
Baseline	Permettez-moi également de dire *en ce moment* que j'ai beaucoup de respect pour les citoyens *du central et oriental - les pays européens* qui, il y a dix ans, ont eu le courage d'entrer dans les rues et de commencer ce processus.
With validated rules	Permettez-moi également de dire *à ce stade* que j'ai beaucoup de respect pour les citoyens *des pays d'Europe centrale et orientale* qui, il y a dix ans, ont eu le courage d'entrer dans les rues et de commencer ce processus.
Source	Clearly, the basic objective of the plan is to stem migration towards the Member States of the European Union and repatriate illegal immigrants living in the Union.
Reference	Il est évident que le but principal de ce plan est de juguler l' émigration vers les pays de l' Union européenne, ainsi que de rapatrier des personnes qui vivent illégalement dans l' Union.
Baseline	Clairement, l'objectif de base du plan est *de refouler la* migration vers les *Etats* membres de l'Union européenne et *de rapatrier* des immigrants illégaux vivant dans l'union.
With validated rules	Clairement, l'objectif de base du plan est *à endiguer* migration vers les *États* membres de l'Union européenne et *rapatrie* des immigrants illégaux vivant dans l'union.

Table 4.10: Examples of improved/regressed translations

Category	**Eng.**	**Fr.**	**Status**
AdjP	Mrs	cher	*discarded*
NP	member	Etat	*discarded*
VP	to have to be	devoir être	*discarded*
NP	NGO	ONG	*retained*
AdvP	in the past	par le passé	*retained*
VP	to be about time	être temps	*retained*

Table 4.11: Examples of discarded/retained entries

4.4 Conclusion

We showed that dictionary extraction could be made effective in improving and customizing a linguistic rule-based system to a specific domain. We described the extraction process and defined an evaluation metric for the quality of dictionary extraction. An error analysis performed on the addition of the extracted rules to the existing, general-purpose system highlighted the various reasons for an ineffective or even damaging application of these new rules. This showed to be mostly due to unwanted interaction with multiple existing rules and, secondly, to the linguistic coding of those entries. In order to avoid regression due to the added dictionary without resorting to a manual inspection, we proposed an automatic, metric-based general solution to select a subset of the extracted rules that would ensure a final improved translation quality. Results on the Europarl domain show an approximately 3 % absolute increase in BLEU on the test set. The work presented in the chapter we are now closing completely relies on an offline processing. It does not change the inner working of the rule-based system. To start with, it does not offer new means of dealing with the ambiguity inherent to the phrasal lexical rules that are being extracted from corpus. This is what the next chapter investigates.

Chapter 5

An Integrated Hybrid System

Based on the qualitative analysis presented in chapter 3, we presented in chapter 4 setups which allow for extraction and offline selection of dictionary entries. In this chapter, we envision a more integrated hybrid system, where not only rules can be extracted from corpora, but statistical decision modules can be used along with the hand-written rules. We still focus on lexical rules only. The offline procedure presented in the previous chapter required the inconvenient choice of a single part-of-speech in the source language and a single meaning in the target language. In addition to that, a unique choice of rule-covering was forced by the hard rule which has longer-spanning entries to always fire over the shorter spanning competing entries. In this chapter, instead of relying solely on an offline filtering of entries (a choice made once and for all inputs), we explore here the possibility of using statistical decision modules to both keep the whole set of entries and decide how to cover the input sentence with the best set of rules.

As a first step in this direction, we start with investigating how much is lost in translation quality when removing a large number of correct entries, due to the lack of a decision module. Oracle experiments artificially look for the best matching with a translation reference. This does of course not provide a real decoding algorithm. It may however inform on the maximum improvement we can expect from decoding.

5.1 Lexical Ambiguity

Even if RBMT systems present a wide diversity in the type of *rules* they use and how they are combined, we can safely assume that they all make use of bilingual dictionaries. While there have been more and more efforts in the field of syntax-based SMT, most SMT systems (Chiang, 2005) do not include any such set of human-readable rules.

Such dictionaries constitute the main entry point in customizing rule-based systems. Workflows for terminology extraction allow to automatically extract such sets of rules from corpora, naturally at the cost of a lower precision than manually created entries.

Ambiguity created by such phrasal (a priori contiguous) dictionary entries is threefold: Part-Of-Speech (POS) tagging, phrasal segmentation and translation. Most of the time, these are handled by separate modules in RBMT systems. Still, as opposed to statistical systems, only one output is kept at each step, until the final translated sentence is produced.

5.2 A look ahead

In the previous chapter, we showed that we could improve translation quality significantly by augmenting it with corpus-extracted phrasal entries. Yet there are a couple of aspects on which we can see room for improvement.

First of all, such lexical rules are extracted with the help of both frequency count heuristics and grammatical constraints that may be improved.

Second, the pruning technique we presented gives an approximate answer to the problem of finding the optimal set of rules. Hill climbing can be used to improve over this stage. Yet the major limitations in the previous work lie in the deterministic aspect of these rules: the direct application of corpus-extracted rules in the deterministic rule-based setup does not allow for ambiguity. We would like to be able to decode the ambiguities generated by the extracted entries instead of both limiting our scope to deterministic rules and having to use an offline pruning strategy.

In order to evaluate an upper-bound over what can be expected from the introduction of decision modules to manage lexical ambiguity, we present oracle experiments. With the artificial hypothesis of a unique answer to the ambiguity in lexical choice, known in advance, we can evaluate the *best* choice we can make using the existing rules, according to the translation quality metric.

The offline pruning algorithm presented in chapter 4 also made use of a reference translation, but on the training corpus only and using a greedy approach in order to make the task of going through that amount of data feasible. Here, oracle experiments directly try to maximize the translation score on the test set, knowing the reference translation. This does not provide with an actual pruning nor decoding scheme, but computationally we can explore a larger space of ambiguities.

5.2.1 Oracle experiment

An *oracle* experiment aims at evaluating an upper bound to the success of a prediction task. In a decoding task, we can distinguish between *model errors* and *search errors*. Model errors are the consequence of lack of data (not enough training material or not fit to the test material or wrong parameters). Whereas search errors are the consequence of the decoding algorithm failing to find the highest scored solution. When the correct prediction is known for the test set, we can look up whether this could have been produced using the model we trained. We may even compute the combination of predictions which would issue the highest score on the test set.

Such results have however to be read with much caution. First of all, because they do not tell which proportion of this can be reached. Also, in the case of Machine Translation, evaluation is done using the proxy of an automatic metric. This type of scoring uses a limited number of (but more than often just one) reference translations. A very good human translation will most certainly not reach a 100% (when relevant, such as for BLEU) score.

Regarding our current problem, we evaluate how much can be expected from predicting the best lexical choice among extracted entries.

We present oracle setups that aim at giving the upper-bound on either pruning the set of deterministic entries or decoding through the whole set of extracted entries. In the first setup, we look for the highest reachable translation score on a given test set, using deterministic lexical translation rules.

5.2.1.1 Oracle for the optimal pruning of corpus-extracted deterministic rules

In order to get an upper-bound of what may be achieved with this method, we apply the previously described rule-filtering algorithm, where we select entries based on the global BLEU improvement or degradation on the relevant test set. Algorithm 2 describes the method we use. It aims at optimizing, though in a greedy fashion, the final score

We start with a set of candidate rules and an empty set of validated rules. We then process to validate rules by batches composed of same source span entries. Therefore, the first batch is composed of one-gram matching entries. The validation corpus is

Algorithm 2 Oracle Algorithm for pruning deterministic entries

```
current set of rules Cur $\leftarrow \emptyset$
Scur $\leftarrow$ 0
Sprev $\leftarrow$ Scur
**for** $i = 1$ to $N$ **do**
    remove all rules of span i from Cur
    **for all** entry $e$ such that $source_span(e) = itokens$ **do**
        compute translation Tref of corpus Ci(Ri) with Cur
        compute translation Tcand of corpus Ci with Cur + Ri
        **if** BLEU(Tcand) $>$ BLEU(Tref) **then**
            add Ri to ValidatedSet(i)
        **end if**
    **end for**
    add ValidatedSet(i) to Cur
    compute translation of devset with Cur
    compute Snew=BLEU(devset)
**end for**
```

indexed by these rules and each rule *e* is thus assigned a subcorpus where it matches a source phrase. This subcorpus Ci(e) is translated two times. First, with the current set of validated rules, and then with the current set augmented by the candidate rule. Translation scores of both translations are then compared. All entries of span N are processed in this way. When all entries have been scored, we proceed to augmenting the set of validated rules with rules of span N. For each rule, if the translation which makes use of the additional rule *e* is scored higher, then the entry *e* will be added to the set of validated rules. Following this, span N+1 can be processed.

This goes on until we have covered rules of span Nmax. We chose to cover up to six-gram spanning entries.

5.2.1.2 Oracle for decoding through the unpruned set of entries

In this setup, we aim at finding an upper-bound of a sentence-level decoding of two ambiguities: partitioning into lexical entries and choice of meaning. The first issue arises when a sequence of words can be covered both by single word matching entries and phrasal entries. These phrases may even overlap. The second issue relates to the multiple choices available to translate the same single word or phrase.

Algorithm 3 Oracle Algorithm with ambiguous entries

```
TextOracle
for all $sentencesS$ do
    trans(S) = SentenceOracle(S)
end for


SentenceOracle(S)
Entries(S) $\leftarrow$ all matching entries
ReferenceTargetStems $\leftarrow$ all stemmed ngrams in reference
Translations(S) $\leftarrow \emptyset$
repeat
    for all $entry\ e$ such that $e \in Entries(S)$ do
        if TargetStem(e) $\notin$ ReferenceTargetStems then
            Remove e from Entries(S)
        end if
    end for
    translation Twith = translate(S,Entries(S))
    add Twith to Translations(S)
until Entries(S)=$\emptyset$


Translate(S,R)
output $\leftarrow$ translation with longest spanning rules
for all entry $e$ such that $e \in longest_spanning(S,R)$ do
    remove e from R
end for
return output
```

In this second setup (Algorithm 3), as opposed to Algorithm 2, we use the complete set of extracted rules and allow two types of ambiguities: partitioning of the source sentence into lexical units and choice of meaning. This experiment intends to evaluate an upper bound for the performance of a rule-based system which would be augmented in the following way. Extracted phrasal rules would be used as alternatives to translate the source sentence, exploring the inherent ambiguity. On that latter aspect, this therefore goes further than the previous experiment, where no disambiguation issue was raised by the addition of the extracted rules. Moreover, this does not say anything about how such a disambiguation would work.

In this oracle experiment, Part-of-Speech disambiguation remains managed by the (rule-based) disambiguation module. The segmentation ambiguity is explored recursively starting from the coverage of the source sentence that uses the fewest long-spanning rules (as it is the default in the rule-based system). A stemmer for the target language allows to discard early choices that do not match the reference sentence. We thus avoid exploring translations with rules whose stemmed target does not match the reference.

5.2.2 Experimental results

5.2.2.1 Oracle pruning of corpus-extracted deterministic rules

In this section, we position the current result on pruning the deterministic phrasal lexical rules with respect to the oracle of section 5.2.1.1. According to Figure 5.1, we would think there is still a reasonable room for improvement in this setup. However, a quick manual inspection (50 sentences) of the oracle translation compared to the current best translation obtained gives 60% degraded translations, for 30% improved and 10% equal. The translation metric is obviously gamed by the Oracle algorithm, resulting in a high score that is not quite correlated with translation quality.

Figure 5.1 shows the current result compared with the oracle described in section 5.2.1.1.

5.2.2.2 Oracle decoding through the unpruned set of entries

For computational reasons, we experiment with a relatively small training set of 60k sentence pairs in the news domain, for the French to English language pair. For this language pair, we compute the sentence-level oracle of the rule-based system augmented

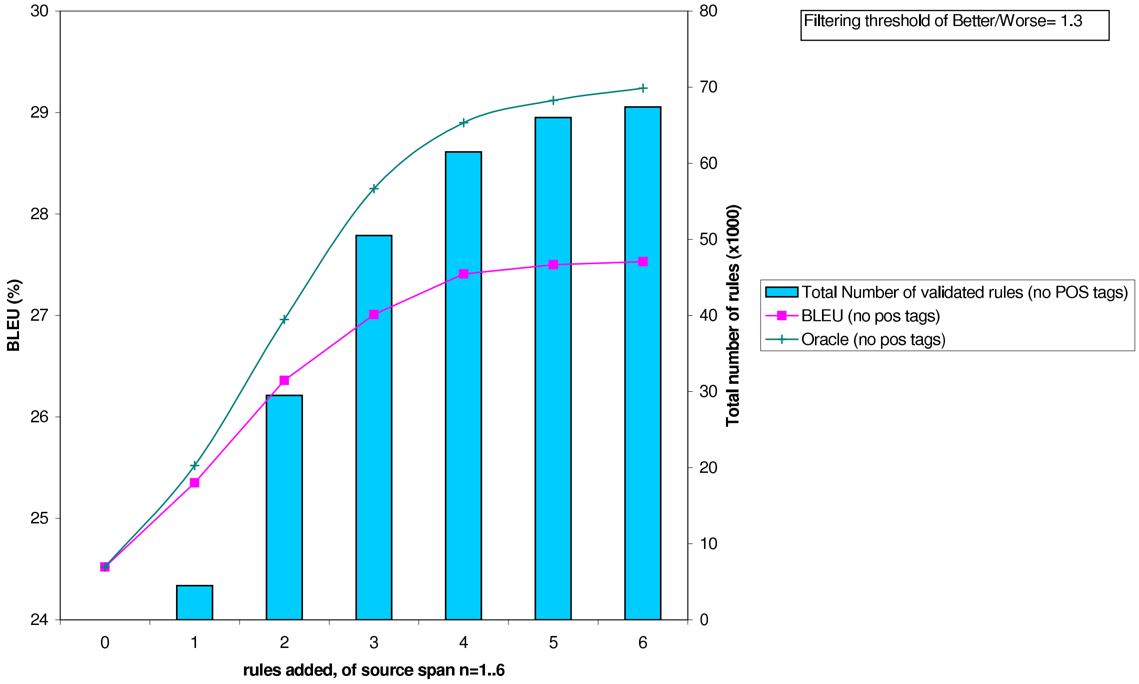


Figure 5.1: Comparison of current algorithm with Oracle (one pass only, no hill-climbing)

with the (unpruned) set of entries. We also compare it with the following alternative systems:

- vanilla rule-based translation
- vanilla Statistical Post Editing translation
- vanilla phrase-based translation
- Oracle for the (text-based) pruned set of extracted deterministic phrasal entries
- Oracle for the (sentence-based) decoding of the unpruned set of phrasal entries

We add to these systems results using the post-editing setup:

- plain post-editing
- a plain post-editing model on the oracle translation
- an Oracle-trained post-editing model on the Oracle translation

Results in Table 5.1 show that the Oracle for the augmented rule-based system is slightly above the vanilla phrase-based result and does not quite reach the score of the statistical-post-editing of the non-augmented rule-based system.

The sole extraction of phrasal entries has no hope to reach the level of the plain phrase-based model, even less the level of the post-editing combination system. Since

#	System	% BLEU
0	phrase-based baseline	26.88 (BP=0.976)
1	original rule-based system	22.69 (BP=1.00)
2	single best meaning validated entries	24.05 (BP=1.00)
3	Oracle pruning of deterministic rules	24.58 (BP=1.00)
4	sentence-level segmentation Oracle: all entries, all segmentations + single best meaning	25.34 (BP=1.00)
5	sentence-level Oracle: all entries, all segmentations, all meanings	27.98 (BP=1.00)
6	raw systran + raw post-editing model	28.29 (BP=1.00)
7	sentence-level Oracle + raw post-editing model	30.50 (BP=0.961)
8	sentence-level Oracle + post-editing model trained on sentence-level Oracle	32.41 (BP=0.968)

Table 5.1: FREN: BLEU scores are tokenized, lowercased BLEU scores on nc-test2007 + 2008, 1556 sentences less than 20 words only. BP= Brevity Penalty

the sentence-level Oracle barely reaches the translation score of the SPE model, we see that such a system might only compete with it if both are used in combination.

5.3 An empirical module for lexical ambiguity in the Rule-Based system

In order not to rely on any specificity of the rule based system, and supposing it is possible to impose choices of lexical entries to the rule-based translation workflow, we propose a preprocessing step that outputs an n-best list of lexical choices. From the source sentence, applicable phrasal entries are collected thanks to the dictionary-matching automata. The covering of all tokens (words) in the source sentence with the matching dictionary entries can be represented by a word lattice. This lattice represents the three types of ambiguity aforementioned. This is illustrated by figure 5.2 for the translation into French of the English sentence “the small blue box fills up with stars”.

The problem of choosing the best set of dictionary entries to cover the source sentence thus modelised can now be expressed as the problem of finding the best path through the lattice.

We propose to use source-side features to weight paths in the lexical ambiguity graph such as represented by Figure 5.2. The default choices made by the rule-based system constitute an important feature that allows the decoding to back off to the rule-based choice.

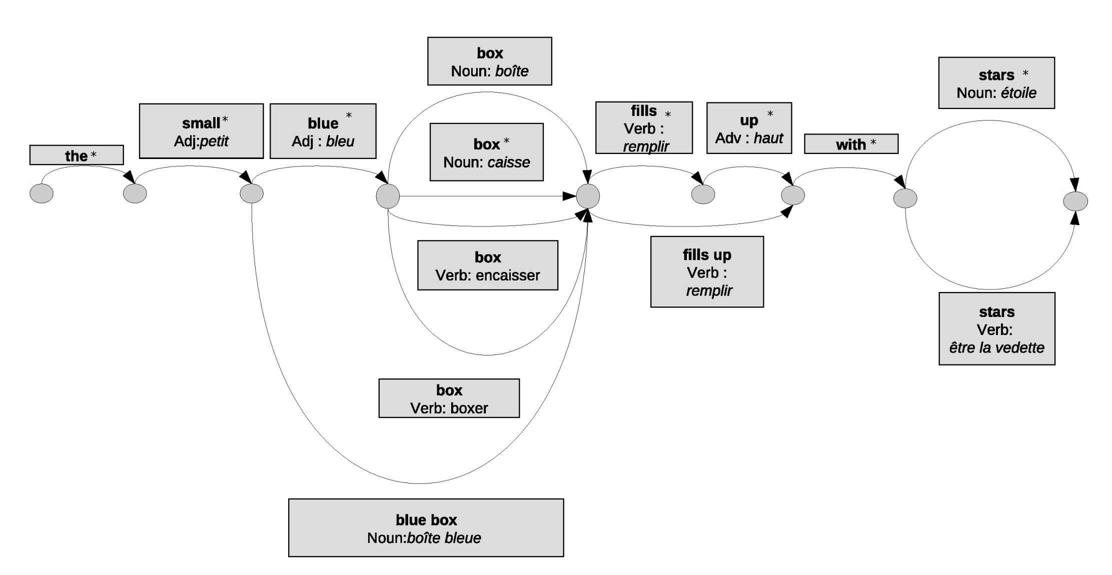


Table 5.2: Lattice representation of dictionary-generated ambiguities. Edges marked with a * are the default rule-based translation options.

From this, we are able to decode an n-best list of lexical choices, before a final synthesis of the target sentence.

This n-best list allows to enforce a limited number of combinations in terms of lexical choices to the generation module. The final target sentences can then be rescored using an n-gram target language model.

5.3.1 Translation Model

From a parallel corpus, additional bilingual phrasal entries can be extracted using a terminology extraction workflow as described in the previous chapter or by Morin et al. (2007). They bear the same features as entries in a phrase-table (Koehn et al., 2007): translation probabilities and lexical weighting in both directions. The source-side lattice of matching rules, initially built upon rule-based choices gets enriched by these additional in-domain entries. Rule-based entries are assigned a uniform, global weight that accounts for the probabilistic features. This will be further on discriminatively tuned.

Let us describe the different models we can use in a log-linear combination to score lexical choices on the source-side lattice.

5.3.2 Hidden Markov Model of Part-Of-Speech tagging

We introduce an n-gram Part-Of-Speech model that aims at choosing dictionary entries in coherence with plausible Part-Of-Speech taggings in the source language. This includes three types of scores. Two edge scores as described in equations 5.1 and 5.2, along with an n-gram model of Part-Of-Speech sequences. We use the decision-tree based *Tree Tagger* (Schmid, 1994) to tag the corpus and thus provide with the training data for the n-gram based model.

$$P(Cat/Ngram) = \frac{\sum Ngram, Cat}{\sum Ngram} \tag{5.1}$$

$$P(Ngram/Cat) = \frac{\sum Cat, Ngram}{\sum Cat} \tag{5.2}$$

5.3.3 Features

We use a 5 gram language model trained on the target side of the parallel corpus to rescore an n-best list of the source lattice decoding.

In addition to the above models, we use the following features on the translation options:

1. default rule-based choice (binary feature that assigns a constant penalty)
   We aim at *augmenting* the existing hand-crafted rules. Our intuition is that most general cases are described by these linguistically informed decision modules. Scoring them specifically may ensure we at least do as well as the Rule-Based choice.

2. target word count penalty
   This is a necessary feature to avoid a bias of the Language Model towards too short translations

3. source phrase count penalty
   This feature aims at balancing the choice between a small number of very lexicalised rules that embed context, and isolated single word phrases that allow for a more flexible recomposition.

5.3.4 Decoding

In the rule-based system we experiment with, disambiguation choices are made in sequence, with little to no possibility of going backward in the chain of decisions. Lexical choices are decided early on. This is why we have to rely on source-side features to find the most suitable path that describes the lexical coverage. Only in a rescoring phase of a limited number of combinations can we resort to target-side models such as an n-gram model.

To perfom this, we use an available Viterbi decoder to decode through the source-side lattice and provide an n-best list of lexical choices. For each of these outputs, choices are enforced relating to segmentation into lexical units, resulting Part-Of-Speech ambiguity and target translation choice for a set segmentation and Part-Of-Speech tagging. The synthesis module of the Rule-Based system is further on in charge of producing the target sentences, theoretically solving inflection and agreement issues.

5.3.5 Discriminative weighting

Since the setup outputs an n-best list of translations of a log-linear model, an optimization algorithm such as Minimum Error Rate Training (Och, 2003) can be used to tune the relative weights of these features towards the evaluation metric of choice.

Model	Setup	Features involved (cumulative)	%BLEU
0	Rule-Based (RB)	(no decoding)	21.8
0bis	Rule-Based (RB)	default choice (side effects)	21.3
1	(0) + Part-Of-Speech ambiguity	default rule, target LM, word count penalty	21.4
2	(1) + P.O.S. model	P(form/POS),P(POSform),P(POS sequence)	21.5
3	(2) +extracted rules	translation probabilities, lexical weights	22.8
5	(3) +10M words source corpus	all	23.0

Table 5.3: Effect of the different features. Scores are computed on nc-test2007 test set(lowercased, tokenized) for French-English.

The only shortcoming here comes from the use of a target language n-gram model in a rescoring phase only: the rule-based process is run a limited number of times on a first n-best list of source-side decoded combinations. Only then, the generated target sentences get rescored by the n-gram target language model.

5.4 Experiments on the augmented system

We experiment on the *news commentary* domain (Callison-Burch et al., 2007), on the French-English language pair. The goal of this experiment is to show we can get a better output by disambiguating even only lexical rules within the rule-based system.

We first want to investigate how the shallow features we mentioned allow a decoding of lexical choices that is independent from the inner working of the rule-based engine. We then want to know how much the rescoring phase necessary to use the target language model impacts on the performance.

We use features that aim at modeling lexical ambiguity along the three aspects we have mentioned. Table 5.3 shows how these features help in obtaining the final performance.

5.4.1 Effect of the size of n-best lists

We do not embed the target sentence generation into the decoding process, both for practical reasons and in order not to be too specific about the Rule-Based system.

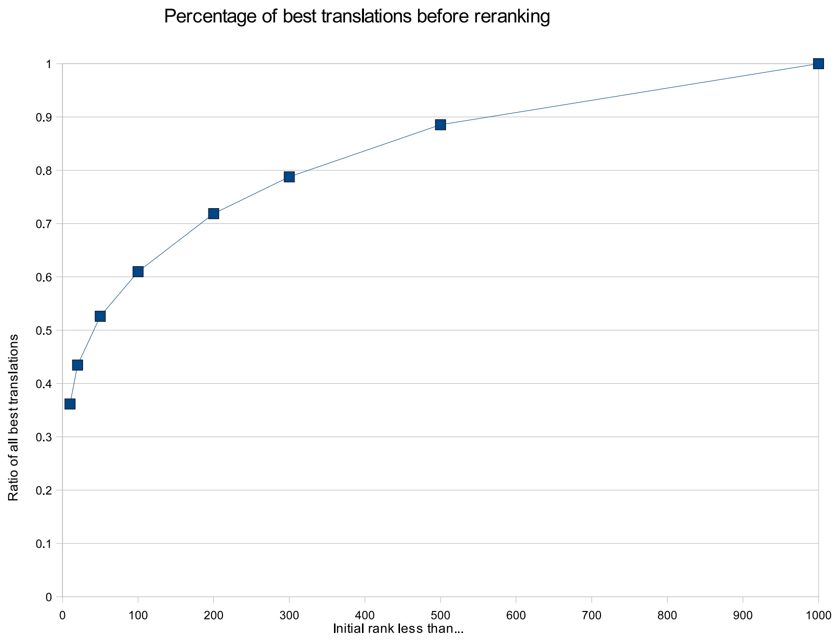


Table 5.4: Effect of reranking, taking 1000 best reranking as reference

Additionally, translation speed is impacted linearly with the size of this list. We experiment here on the impact of the n-best list size on translation quality.

Both training and testing with an n-best list size of 1000 (this is the number of distinct paths in the source side lattice, which do not necessarily lead to distinct translation outputs), we evaluate the ratio of one-best translations captured when reducing the size of the list to be rescored. On Figure 5.4, we see for example that it takes an n-best list of 200 to capture 90% of the best outputs computed with a 1000-best list.

Table 5.5 displays the translation scores according to the size of the rescored n-best list.

This shows that translation quality is strongly impacted by not integrating the language model in the search algorithm.

5.4.2 Results

A manual comparative evaluation on a test set of 100 sentences (Table 5.6) shows that, though scoring higher on an evaluation metric (+5 BLEU points), a phrase-based model is evaluated as of lower overall quality on a sentence-by-sentence basis than the

Size of rescored N-best	**%BLEU**
1	21.35
10	20.51
20	21.47
50	22.19
100	22.95
500	23.36
1000	23.38

Table 5.5: Effect of the size of the n-best list on translation quality (nc-test2007, fren)

System	SMT	RBMT	RBMT + extracted entries
SMT	-	0.3	0.3
RBMT	0.6	-	0.3
RBMT +	0.6	0.6	-

Table 5.6: Manual evaluation : ratio of sentences judged better for system A (row) than system B (column)

rule-based system we experiment with. Moreover, augmenting the rule-based system with extracted dictionary entries on the same training data significantly improves on the rule-based baseline.

5.5 Conclusion

Beyond the efficiency aspect, the experiments we conducted show a gain in combining the most general rule-based structure with a corpus-based extraction of lexical rules. Once again, manual evaluation contradicts the results of automatic metrics. A manual inspection would lead to conclude that a corpus-based extraction of rules whose structure is linguistically constrained (here, phrasal dictionary entries) is superior to both a non-customized rule-based system and a statistical system based on the same data. This while the comparison of automatic scores shows little difference between hybrid and purely statistical phrase-based systems.

Chapter 6

Conclusion and perspectives

We introduced this work with a very broad intention of exploring the convergence of years-long diverging approaches to the problem of Machine Translation. For this sake, we had to give in to an exercise of clarification of what exactly was meant by "Rule-Based" and "Statistical" Machine Translation. Following this, we made an effort into describing a particular such *RBMT* system, in spite of many intricacies that characterize such architectures. We also described a vanilla *Phrase Based* Machine Translation system. Armed with the definitions of those terms and an overview of the specific systems, we went on with a review of the relevant work in system combination or so-called *hybrid* systems. It appeared that very few attempts at this point had been made to combine *RBMT* and *SMT* approaches. We designed initial "black box" architectures and included them in a pool of systems trained on the same dataset. This set the basis for a comparative study of rule-based, statistical and combined architectures. Comparison showed that a hybrid system could be made to surpass both approaches in isolation. And we could draw qualitative lessons in terms of a prospective hybrid solution, differentiating the respective benefits of both approaches in terms of word order, terminology and choice of grammatical words. The performance of those systems built on the same dataset could reveal differences that showed complementarity between the hand-crafted, linguistics-driven rule-based approach and the empirical models and decoders.

We then went on to experimenting with one aspect of hybridizing: automatic extraction of translation rules from corpora. On the restricted scope of lexical phrasal rules, we describe a learning pipeline which produces linguistically encoded dictionary entries. We also provide a means to prune the initial set of extracted entries, so that translation quality can be maximised, as approximated by an evaluation metric.

This latter experiment combines the still linguistically structured rule-based architecture with both corpus-extracted rules and a discriminative pruning through these rules.

We made a last attempt at embedding yet another feature of statistical models into a rule-based architecture, by exploring empirical models for disambiguation. Starting with Oracle models, we then experiment with a log-linear combination of models that aim at solving three types of ambiguity (Part-Of-Speech tagging, phrasal segmentation and translation choice) altogether. We show such a system can be effective, however limited by the inherent structure of the rule-based architecture.

To go further into the directions we started to explore, we would identify a room for improvement in the quality of the extracted rules. Syntactic rules could be learnt that go further than the scope of phrasal lexical rules. Discriminative pruning of rules could be extended to manually input rules: lexicographers could suggest rules that then would get scored by statistical techniques, so as to maximize translation quality on a held-out corpus. Much more could be done in the realm of combining both manually input disambiguation rules and soft decision models such as n-gram language models.

Finally, we have shown that rule-based and statistical approaches displayed complementary qualities. Depending on the language pair and the domain, it may be relevant to choose one or the other. However costly, it is still necessary to perform manual evaluation, especially if automatic metrics scores remain close. Because of some specific errors in long-range word order or crucial grammatical words, a rule-based system may still be prefered.

A second lesson learnt is that combining both approaches can be effective in ensuring to benefit from both an existing rule-based system and available data. The most straight-forward and massively impacting method when wanting to keep the rule-based architecture is the extraction of dictionary entries.

The third lesson learnt in this study is that for a more thorough combination of rule-based and statistical approach, the exisitng rule-based implementation may be crucial in helping or impeding the incorporation of statistical decision modules. Ideally, to experiment further in that direction, it would be useful to design the input of both manual and corpus-extracted rules in the translation engine.

Bibliography

AbduI-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve smt performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23. Association for Computational Linguistics.

Alex, B. (2002). Using language models to assist in the correction of machine translation output. Master's thesis, School of Informatics, University of Edinburgh.

Auli, M., Lopez, A., Hoang, H., and Koehn, P. (2009). A systematic analysis of translation model search spaces. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 224–232, Athens, Greece. Association for Computational Linguistics.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, ACL*.

Bangalore, S., Bordel, G., and Riccardi, G. (2001). Computing consensus translation from multiple machine translation systems. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001)*.

Bannard, C. (2006). *Acquiring Phrasal Lexicons from Corpora*. PhD thesis, School of Informatics, University of Edinburgh.

Birch, A., Blunsom, P., and Osborne, M. (2009). A quantitative analysis of reordering phenomena. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 197–205, Athens, Greece. Association for Computational Linguistics.

Bod, R. (1992). A computational model of language performance: Data oriented parsing. In *Proceedings of the 14th conference on Computational linguistics-Volume 3*, pages 855–859. Association for Computational Linguistics Morristown, NJ, USA.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, page 116. Association for Computational Linguistics.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., and Roossin, P. (1988). A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics*, pages 71–76, Morristown, NJ, USA. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *WMT workshop*. ACL 2007.

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, Columbus, Ohio. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *EACL 2006*.

Carbonell, J. G., Probst, K., Peterson, E., Monson, C., Lavie, A., Brown, R. D., and Levin, L. S. (2002). Automatic rule learning for resource-limited mt. In *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pages 1–10, London, UK. Springer-Verlag.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

Chomsky, N. (1957). *Syntactic structures. The Hague: Mouton.. 1965. Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: The MIT Press.

Cicekli, I. and Guvenir, H. (1996). Learning translation rules from a bilingual corpus. In *Proceedings of NEMLAP-2*.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause restructuring for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540, Morristown, NJ, USA. Association for Computational Linguistics.

Daille, B., Gaussier, E., and Lange, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of COLING 94*.

DeNeefe, S., Knight, K., and Chan, H. H. (2005). Interactively exploring a machine translation model. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 97–100, Ann Arbor, Michigan. Association for Computational Linguistics.

DeNeefe, S., Knight, K., Wang, W., and Marcu, D. (2007). What can syntax-based MT learn from phrase-based MT? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 755–763.

Dugast, L., Senellart, J., and Koehn, P. (2007). Statistical post-editing on SYSTRAN's rule-based translation system. In *WMT*. ACL 2007.

Dugast, L., Senellart, J., and Koehn, P. (2008). Can we relearn an rbmt system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, U.S.A. Association for Computational Linguistics.

Dugast, L., Senellart, J., and Koehn, P. (2009a). Selective addition of corpus-extracted phrasal lexical rules to a rulebased machine translation system. In *Proceedings of Machine Translation Summit XII*, pages 222–229.

Dugast, L., Senellart, J., and Koehn, P. (2009b). Statistical post editing and dictionary extraction: Systran/Edinburgh submissions for ACL-WMT2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 110–114, Athens, Greece. Association for Computational Linguistics.

Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., and Chen, Y. (2008). Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *WMT workshop*. ACL 2008.

Fox, H. J. (2002). Phrasal cohesion and statistical machine translation. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 304–3111, Morristown, NJ, USA. Association for Computational Linguistics.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA. Association for Computational Linguistics.

Hogan, C. and Frederking, R. E. (1998). An evaluation of the multi-engine mt architecture. In *AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 113–123, London, UK. Springer-Verlag.

Hu, X., Wang, H., and Wu, H. (2007). Using RBMT systems to produce bilingual corpus for SMT. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 287–295.

Imamura, K., Sumita, E., and Matsumoto, Y. (2003). Feedback cleaning of machine translation rules using automatic evaluation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 447–454, Morristown, NJ, USA. Association for Computational Linguistics.

Itagaki, M., Takako, A., and He, X. (2007). Automatic validation of terminology translation consistency with statistical method. In *Proceedings of MT Summit XI*.

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Comput. Linguist.*, 25(4):607–615.

Knight, K. and Chander, I. (1994). Automated post-editing of documents. *Proceedings of the Twelfth National Conference on Artificial Intelligense*, pages 779–784.

Koehn, P. (2003). *Noun phrase translation*. PhD thesis, USC. Adviser-Kevin Knight.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT summit 2005*.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge MIT Press.

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)*, pages 868–876.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL 2007, demonstration session*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Kumano, A. and Hirakawa, H. (1994). Building an mt dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*. ACL 1994.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *31st Annual Meeting of the Association for Computational Linguistics*. ACL 1993.

Langkilde, I. and Knight, K. (1998). The practical value of N-grams in derivation. In Hovy, E., editor, *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 248–255, New Brunswick, New Jersey. Association for Computational Linguistics.

Lavie, A. (2008). Stat-xfer: A general search-based syntax-driven framework for machine translation. In *Proceedings of CICLing-2008*, pages 362–375.

Llitjos, A. F. and Carbonell, J. G. (2006). Automating post-editing to improve mt systems. In *Automated Post-Editing Workshop at AMTA*.

Llitjos, A. F. and Vogel, S. (2007). A walk on the other side: Adding SMT Components in a Transfer-Based Translation System. In *SSST workshop*. NAACL-HLT 2007 / AMTA.

Macherey, W. and Och, F. J. (2007). An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007*

Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 986–995.

Manning, C. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Menezes, A. and Richardson, S. D. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of ACL 2001*.

Moore, R. C. (2001). Towards a simple and accurate statistical approach to learning translation relationships among words. In *Proceedings of the workshop on Data-driven methods in machine translation*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671, Prague, Czech Republic. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of ACL-2003*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *ACL '02*. ACL 2002.

Popovic, M. and Ney, H. (2007). Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55, Prague, Czech Republic. Association for Computational Linguistics.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency treelet translation: syntactically informed phrasal smt. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 271–279, Morristown, NJ, USA. Association for Computational Linguistics.

Ramırez-Sánchez, G., Sánchez-Martınez, F., Ortiz-Rojas, S., Pérez-Ortiz, J., and Forcada, M. (2006). Opentrad apertium open-source machine translation system: an opportunity for business and research. In *Proceedings of the Twenty-Eighth International Conference on Translating and the Computer*. Citeseer.

Richardson, S. D., Dolan, W. B., Menezes, A., and Pinckham, J. (2001). Achieving commercial-quality translation with example-based methods. In *Proceedings of MT summit 2001*.

Rosti, A.-V., Ayan, N. F., Xiang, B., Matsoukas, S., Schwartz, R., and Dorr, B. (2007). Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York. Association for Computational Linguistics.

Sadat, F., Johnson, H., Agbago, A., Foster, G., Kuhn, R., Martin, J., and Tikuisis, A. (2005). PORTAGE: A phrase-based machine translation system. *Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, 100:129.

Salkoff, M. (1999). *A French-English grammar: a contrastive grammar on translational principles*. John Benjamins Publishing Company.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Senellart, J., Yang, J., and Rebollo, A. (2003). Systran intuitive coding technology. In *in Proceedings of MT Summit IX*.

Simard, M., Goutte, C., and Isabelle, P. (2007a). Statistical phrase-based post-editing. In *NAACL-HLT 2007*.

Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. (2007b). Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 203–206, Prague, Czech Republic. Association for Computational Linguistics.

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*.

Thurmair, G. (2004). Comparing rule-based and statistical mt output. In *Proceedings of the Workshop on the amazing utility of parallel and comparable corpora, LREC*.

Thurmair, G. (2005). Hybrid architectures for machine translation systems. *Language resources and evaluation*, 39:91–108.

Toma, P. (1972). Optimization of systran system. Technical report, LATSEC Incorporated.

Ueffing, N., Stephan, J., Matusov, E., Dugast, L., Foster, G., Kuhn, R., Senellart, J., and Yang, J. (2008). Tighter integration of rule-based and statistical MT in serial system combination. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 913–920, Manchester, UK. Coling 2008 Organizing Committee.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Machine Translation Output. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702, Genoa, Italy.

White, J. S. (1985). Characteristics of the metal machine translation system at production stage. In *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23(3):377–403.

Wu, H. and Wang, H. (2009). Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 154–162. Association for Computational Linguistics.

Xia, F. and McCord, M. (2004). Improving a statistical mt system with automatically learned rewrite patterns. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 508, Morristown, NJ, USA. Association for Computational Linguistics.

Yamada, K. and Knight, K. (2001). A decoder for syntax-based statistical mt. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational*

Linguistics, pages 303–310, Morristown, NJ, USA. Association for Computational Linguistics.

Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City. Association for Computational Linguistics.

Zollmann, A., Venugopal, A., Och, F., and Ponte, J. (2008). A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1145–1152, Manchester, UK. Coling 2008 Organizing Committee.