Selective addition of corpus-extracted phrasal lexical rules to a rule-based machine translation system

Loic Dugast[1,2] and **Jean Senellart**[1]
[1]SYSTRAN S.A.
La Grande Arche
1, Parvis de la Défense
92044 Paris
La Défense Cedex
France

Philipp Koehn[2]
[2]University of Edinburgh
10 Crichton Street,
Edinburgh
United Kingdom

Abstract

In this work, we show how an existing rule-based, general-purpose machine translation system may be improved and adapted automatically to a given domain, whenever parallel corpora are available. We perform this adaptation by extracting dictionary entries from the parallel data. From this initial set, the application of these rules is tested against the baseline performance. Rules are then pruned depending on sentence-level improvements and deteriorations, as evaluated by an automatic string-based metric. Experiments using the Europarl dataset show a 3% absolute improvement in BLEU over the original rule-based system.

1 Introduction

Rule-based systems generally make use of manually written structural transfer rules. They also use a greater number of lexical rules, most often designated as *dictionary entries*, as can be understood in the general sense of human-purpose bilingual dictionaries. They contain word and phrasal entries of various categories. Such dictionaries are the easiest and most frequently modified components of such systems. Historically these dictionaries have been built by linguists who have manually entered source words and their translations along with the corresponding linguistic information: part of speech, head word, inflection type. User interfaces have often been developed to enable quick and mostly automatic linguistic coding of such rules. On the contrary many statistical systems which have emerged claim to get high coverage through extracting all these mappings from parallel corpora, often without any linguistic constraint. Regarding domain, rule-based systems have most of the time been designed to be general-purpose, requiring a customization effort to adapt to a specific topic. In this paper, we experiment on English to French translation in the Europarl domain and add a set of 67,000 dictionary entries to the general-purpose SYSTRAN rule-based system which translate into a 3 %BLEU absolute improvement in translation quality.

1.1 Motivation

Let us first explain what motivates a need for bilingual phrase dictionaries. This motivation is twofold. Previous experiments indicated that most of the improvements of a statistical phrase-based layer in a combination with a rule-based system came from lexical changes, most of them phrasal expressions. Then, the distribution of bilingual phrases in terms of phrase length and the nature and availability of manually written phrasal bilingual dictionaries for a given domain is an incentive for corpus extraction.

The first argument is illustrated both by a combination of different rule-based systems (Eisele et al., 2008) and a statistical post-editing layer over a rule-based output (Simard et al., 2007), along with some qualitative analysis (Dugast et al., 2007).

As for the second argument, in the past few years, statistical machine translation moved from word-based models to phrase-based[1] models. In a more linguistic approach, Bannard (2006) discusses the

[1] Here, the word ”phrase” simply denotes any sequence of words, not necessarily a constituent.

	English	French
1	big park	grand parc
2	private bank	banque privée
3	left bank	rive gauche
4	fig leaf	feuille de vigne
5	fraud scandal	scandale en matière de fraude
6	freight traffic	traffic de marchandises
7	to let off steam	décompresser

Table 1: Examples of phrasal entries

existence of a phrasal[2] lexicon and mentions syntactic fixedness and lack of compositionality as hints for its existence.

We see three main reasons to want to learn phrasal entries for the rule-based system. A first reason lies in the ability to capture local context to disambiguate the translation (as in examples 1-2 of Table 1). Then, there are phrases that cannot be translated word-for-word, such as examples 4 and 5. And finally, some strong collocations may reduce the syntactic ambiguity of the source sentence (examples 6 and 7). We however probably do not need entries such as *big park* in Table 1, for its translation into French is compositional, little ambiguous and the English phrase *big park* could be easily modified into *very big park*, *big natural park*.

1.2 Related work

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, one variation between the different approaches lies in the choice of either extracting all monolingual terms to then find alignments or align chunks of raw text (typically, extract a phrase table) which is then filtered to keep only the syntactically meaningful ones. Daille (1994) and Kumano (1994) belong to the first category, while Itagaki (2007) belongs to the second category. Another variation is the choice of confidence measures to evaluate the quality of a candidate entry. As far as application of such dictionaries is concerned, Melamed (1997) uses mutual information as the objective function maximized during the learning process. Font Llitjos et al. (2007) describes a semi-automatic procedure to extract new dictionary entries in a rule-based system. This however deals with a small number of entries and necessitates a manual review. Itagaki (2007) presents a filtering method for candidate entries using a Gaussian Mixture Model classifier trained on human judgements of the quality of a dictionary entry, but does not provide evaluation of final translation quality. The closest work from what we describe in the present paper might be the one by Imamura (2003), in which example-based pattern rules are filtered using an automatic evaluation of the final translation output. In the work presented here, we describe two independent training steps that first extract dictionary candidates and then automatically validate them directly within the RBMT system.

2 Dictionary extraction

2.1 Manual coding of entries

The SYSTRAN rule-based system provides a dictionary coding tool (Senellart et al., 2003) that allows the manual task of coding entries to be partially automated thanks to the use of monolingual dictionaries (Table 2), morphological guess rules and probabilistic context-free local grammars (Table 3). For example, the second rule illustrated in the latter table simply describes how an English noun phrase may be composed of a *adjective+noun* sequence. The general rule has a phrase to inherit inflection and semantic (we won't mention this aspect here, as it is of little significance) features from the headword. The coding tool also allows the user to fine-tune it by correcting the automatic coding and/or add more features. However, this remains a time-consuming task. Moreover, it is not easy for humans to select the best translation among a set of alternatives, let alone assign them probabilities.

2.2 Extraction from corpus

The extraction setup as depicted in Figure 1 starts from a parallel corpus dataset. A state of the art procedure is followed (word alignment using GIZA++ in both directions and use of heuristics to extract phrase pairs) to extract a phrase table.

Each phrase pair is then processed by the dic-

[2] This time, the word "phrase" is understood as a syntactic constituent.

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 2: Sample of the monolingual dictionary for English

rule	headword index	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 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

tionary coding engine. We restrict the extraction to entries for which the target syntactic category is identical with the source category. In that concern, Koehn (2003) evaluated 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. As a bilingual entry may have various inflectional forms in the corpus, we then have to sum the lemma counts. Only then are the frequencies relevant to perform filtering on these entries. This will be used to retain only one translation per source phrase, either the most frequent or the best aligned (according to lexical weights) in the case of multiple highest frequency translations.

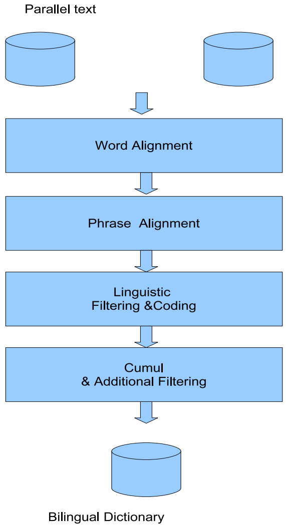

Figure 1: Extraction pipeline: from parallel texts to bilingual dictionary

3 Use of the extracted dictionary within the rule-based system

After trying to extract the best and greatest part of all possible good entries (the syntactically relevant phrase pairs) from a parallel corpus, we now examine the issue of using them to improve translation within the rule-based system.

Let us first quickly present the translation flow of the SYSTRAN rule-based system we experiment with. Compound dictionaries are applied after tokenisation. As a default, longest spanning rules match over shorter ones. Then the part-of-speech disambiguation module selects the category whenever multiple rules of different categories may match. Finally, in this setup, only one translation per source is retained for a given category. Since the source words are matched thanks to a finite state automaton of all inflected forms of the dictionary entries, they are mapped to both their lemma and all the possible inflection attributes corresponding to the specific inflected form met in the source sentence. Analysis then uses this morphological tagging to produce a dependency structure. The last stage consists in creating a target language syntactic tree from the isomorphic target dependency tree.

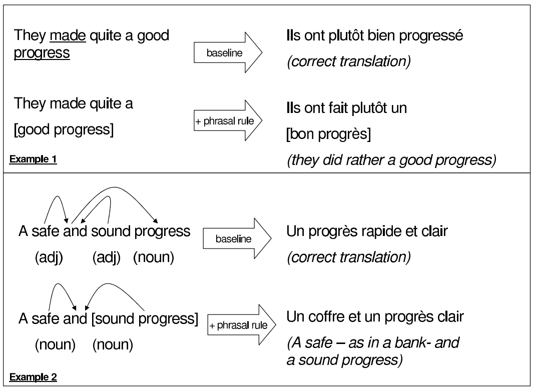
They made quite a good progress
baseline
Ils ont plutôt bien progressé
(correct translation)
They made quite a [good progress]
+ phrasal rule
Ils ont fait plutôt un [bon progrès]
(they did rather a good progress)
Example 1
A safe and sound progress
(adj) (adj) (noun)
baseline
Un progrès rapide et clair
(correct translation)
A safe and [sound progress]
(noun) (noun)
+ phrasal rule
Un coffre et un progrès clair
(A safe – as in a bank- and a sound progress)
Example 2

Figure 2: Examples of deterioration when adding phrasal rules

This synthesis stage may insert function words such as determiners and prepositions. It also generates the inflected form according to the target tree.

4 Validation of entries

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.

Thus, such a verbal expression as *make good progress* that may have been correctly translated would then be mistranslated once the phrasal entry *good progress* is added to the rules' base. A noun phrase such as ”rapid and sound progress” may also get mistranslated from adding *sound progress* as a contiguous noun phrase, as illustrated on figure 2.

Therefore the problem consists of building the optimal subset from the set of candidate entries, according to a translation evaluation metric (here, BLEU (Papineni et al., 2002)), while being constrained by the deterministic firing of these rules.

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. BLEU score measured on a held-out development set is expected to increase at each iteration of adding longer-spanning rules, and to stabilize for the longest spanning entries (here, 6 grams).

Algorithm 1 Dictionary Validation Algorithm

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

5 Evaluation and error analysis strategy

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

5.1 Evaluation of dictionary extraction

The criteria for an entry to be correct are as follows: both word sequences must be phrases of the coded category, the lemma and inflectional codes have to be correct and finally the target phrase must of course 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 good extracted entries among the good extractable entries in the data). Recall especially matters since such a setup for automatic extraction of entries is motivated by its ability to alleviate the human burden of entering entries.

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 interfaces for Translation Memory review and dictionary coding are used for this purpose. This constitutes the Gold Standard dictionary.

In the training process, back pointers to the previous step allows to extract the subset relevant to the Gold Standard training set out of the total extracted dictionary. Precision, recall and consequently F-Measure can then be computed by comparing this extracted subset with the Gold Standard dictionary. This of course makes the assumption that all entries in the Gold Standard are good entries and all good entries we can possibly extract are contained in this Gold Standard dictionary.

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

- alignment: one or both sides of the entry have been truncated
- syntactic category: a verb phrase has been wrongly parsed as a noun phrase, for example
- coding: lemmatisation or identification of the headword is wrong

5.2 Evaluation of translation

Given a certain quality of dictionary, we now face the question of how much translation quality benefits from these dictionaries. 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 of so different architecture 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.

6 Experiments and results

We describe here experiments for both the dictionary extraction and the translation aspects.

6.1 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. If 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 validating. The progress of translation quality as rules are added is evaluated 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. 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, still before any pruning, for each syntactic category by manual judgement on a random sample (Table 5). The 64% precision for noun phrases when using the part-of-speech tags is similar to the result obtained by Itagaki (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 align-

Setup	Precision	Recall	F1
baseline	32%	65%	41%
+ p.o.s.	46%	49%	45%
+validation	52%	n.a.	n.a.
+ p.o.s.+validation	71%	n.a.	n.a.

Table 4: 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 5: Human Evaluation of dictionary extraction (most frequent meaning only)

ment (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 occurs when both segments are aligned and were both identified as phrases of a given category, but are actually truncated parts of a larger bilingual unit, or occasionally when this could have been coded with a different category. Entry #2 of Table 6 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 headword ic crucial because, as a default rule, the entry will inherit its properties from it. Also, only this headword and its identified modifiers in the source phrase will be inflected. This consequently matters for the sake of coverage of inflected source phrases. It also impacts the target side for the sake of generating the correct inflected target phrase.

#	Err. type	English	French
1	Align.	*correction* in the stock	*correction* des bourses
2	Cat.	*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 6: 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 7: 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 8: 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+selected rules

6.2 Application of extracted dictionaries

The baseline system consists in the current rule-based system. System #2 consists in adding the extracted dictionary before validation, while the third system uses only validated rules.

Table 7 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 deal with either part-of-speech ambiguity, regressive interaction with the dependency analysis, and the lack of a mechanism for translation choice or interaction with the existing set of rules.

Figure 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 re-

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 9: Examples of discarded/retained entries

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.
System3	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.
System3	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 10: Examples of improved/regressed translations

call (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 150,000 word entries and a similarly sized phrase dictionary.

Table 8 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 9 illustrates discarded and retained entries while Table 10 shows two samples of compared translations.

7 Discussion

We showed that dictionary extraction could be made efficient 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. Error analysis on the addition of the extracted rules to the existing, general-purpose system highlighted the various reasons for an inefficient or even damaging application of these new rules. 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.

There are margins of progress in both steps of the process. As far as rule extraction is concerned, we may want to try to learn mappings of treelets of the source dependency analysis to target dependency treelets, instead of the sole contiguous phrasal rules (Quirk et al., 2005). The validation process could be improved by replacing arbitrary decisions such as the cut-off threshold and the minimum frequency of validations of a rule by an optimizing step on a small held-out tuning set. Overfitting could be better dealt with than with the sole discarding of singleton entries.

Finally, we recall that alternative translations have been discarded here. This may prevent from validating good novel translations because of the dispersion of the translation of a given source phrase. We would hope also to make the best out of the new extracted rules by decoding among local alternative translations. We intend to address these issues in future work.

References

Colin Bannard. 2006. *Acquiring Phrasal Lexicons from Corpora*. Ph.D. thesis, School of Informatics, Univer-

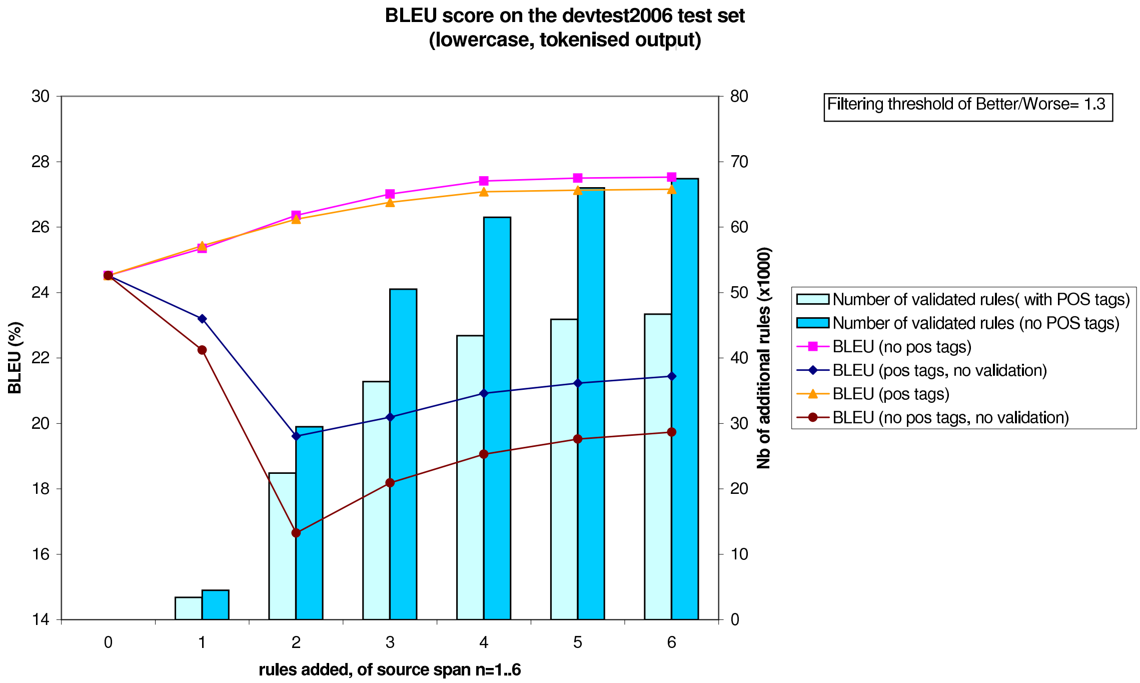

Figure 3: Progress of BLEU score at each iteration of the validation process

sity of Edinburgh.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *EACL 2006*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *WMT workshop*. ACL 2007.

Béatrice Daille, Eric Gaussier, and Jean-Marc Lange. 1994. Towards automatic extraction of monolingual and bilingual terminology.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. In *WMT*. ACL 2007.

Andreas Eisele, Christian Federmann, Hervé Saint-Amand, Michael Jellinghaus, Teresa Herrmann, and Yu Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *WMT workshop*. ACL 2008.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 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.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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.

Philipp Koehn. 2003. *Noun phrase translation*. Ph.D. thesis, USC. Adviser-Kevin Knight.

Akira Kumano and Hideki Hirakawa. 1994. Building an mt dictionary from parallel texts based on linguistic and statistical information. ACL 1994.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. ACL 1993.

Ariadna Font Llitjos and Stephan Vogel. 2007. A walk on the other side: Using SMT Components in a Transfer-Based Translation System. In *SSST workshop*. NAACL-HLT 2007 / AMTA.

Xiaodong He Masaki Itagaki, Takako Aikawa. 2007. Automatic validation of terminology translation consistency with statistical method. In *Proceedings of MT Summit XI*.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. *Proceedings of ACL 1997 workshops*, pages 97–108.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL '02*. ACL 2002.

Chris Quirk, Arul Menezes, and Colin Cherry. 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.

Jean Senellart, Jin Yang, and Anabel Rebollo. 2003. Technologie systran intuitive coding. In *in Proceedings of MT Summit IX*.

Michel Simard, Cyrille Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *NAACL-HLT 2007*.