Rapid Development of a French–English Transfer System

Greg Hanneman
11-731: Machine Translation
Term Project

May 7, 2007

1 Introduction

A key concern in building transfer or rule-based machine translation (MT) systems is the amount of human labor that must be spent writing the necessary bilingual lexicon and transfer grammar. Well-known rule-based systems from past decades (e.g. Systran) were constructed manually over a period of several years, but more recent progress and development has put more emphasis on data-driven statistical techniques. Therefore, an interesting current avenue of research is to explore to what point automatic tools and a more learning-based approach can be used in the development process of a rule-based engine to make system prototyping faster.

The AVENUE project, for example, is based on a "stat-transfer" framework, as described by Peterson (2002), that combines a traditional rule-based transfer MT system with a statistical decoder. Bilingual lexical entries and a transfer grammar with feature unification constraints are applied to the source-language input, and target-language output is synchronously generated as the source is parsed. Possible translations for each parsed structure are stored in a lattice. The final lattice for a sentence is passed to a decoder, which selects the best path through the lattice based on statistical language model probabilities and other parameters. The framework also allows definition of both lexical and rule probabilities, which will also be taken into account as decoding parameters.

Researchers have also considered focusing their development efforts on "subtasks" within MT in the hopes of getting the best results from a reduced amount of labor. There is evidence that the correct translation of noun phrases (NPs) is of particular importance for the success of an overall MT system, and that the subtask of NP translation generalizes well across languages. In a German–English corpus of 100 sentences taken from the proceedings of the European Parliament, Koehn (2003) found that 122 of 168 German NPs had English translations that were also NPs, and furthermore that 164 of the 168 (97.6 percent) *could* be translated as English NPs in acceptable translations of the same sentences. A similar situation was found for Portugese–English and Chinese–English (Koehn and Knight, 2003).

The goal of this project is to invesitage both of these research directions: the introduction of statistical techniques in a rule-based engine, and the importance of noun phrase translation. To address the first, this project will take advantage of the AVENUE framework and other automatic or statistical MT tools to quickly develop a broad-coverage and high-quality French-to-English transfer system with a minimal amount of manual labor. For the second, the usefulness of noun phrase translation as a subtask in system development to improve overall translation quality will also be explored.

2 System Development

Beginning from a training corpus of parallel data, the development work for this project was broken down into five stages: (1) preprocessing the corpus, (2) extracting word-level alignments from it, (3) building a word-level bilingual lexicon, (4) building a phrase-level bilingual lexicon for NPs, and (5) writing a transfer grammar. The following subsections discuss each of these processes individually.

2.1 Corpus Processing

Most of the training data for the system came from Release 3 of the Europarl French–English parallel corpus (Koehn, 2005), representing transcripts of the proceedings of the European Parliament for the years 1996 through 2006. The Europarl corpus is freely available online in 11 European languages[1]; the new Release 3 was prepared especially for the 2007 shared task of the ACL Workshop on Statistical Machine Translation[2].

The corpus is generally aligned by sentence or short paragraph, with one sentence or paragraph per line in both English and French texts. Inequalities in translation length are padded out by the insertion of blank lines when necessary, although some seem to have been inserted incorrectly. Previous releases of the Europarl corpus are also annotated with HTML tags indicating speaker identifications and paragraph breaks.

In addition to the Europarl data, the ACL workshop provided a small amount of "out-of-domain" data taken from a news commentary corpus of editorial-style writing. This also became part of the system training data.

Both halves of the combined parallel corpus were preprocessed to regularize the text to lowercase. Furthermore, when a blank line appeared in the text of either language, the corresponding line in the other language was also discarded. The tokenization on the English side of the corpus was left intact, but additional resegmentation was applied on the French text to recombine apostrophes with the word immediately preceeding them. French apostrophes fulfill much the same role as their English counterparts, indicating missing letters generally at the end of a word, so the retokenization in effect treats tokens like *qu'* and *c'* as different surface forms of *que* and *ce* rather than as bigrams. One exception to the tokenization rule is the French word *aujourd'hui* ("today"), which is lexically and semantically considered one unit. It is therefore left as one token under this system's segmentation scheme.

After processing, the training set comprised 37.2 million words of English running text and 39.2 million words of French running text, divided into more than 1.3 million aligned sentences.

2.2 Word Alignment

Word alignments were extracted from the processed corpus using the GIZA++ alignment toolkit (Och and Ney, 2003) trained to IBM Model 3. Alignments were computed in both the French-to-English and English-to-French directions, and the intersection of these two sets was extracted. This step was intended to remove lower-quality alignments that were not hypothesized independently by both directional alignment processes, but it also has the negative side effect that only one-to-one word alignments are preserved. The final output of the extraction step consisted of a French vocabulary list with English alternatives for each word and a count of the alignment frequency for each pair.

As Figure 1 shows, the French–English alignments are still rather noisy. Therefore, the possible English alternatives for each French word are further filtered based on their frequency counts in order to remove infrequent, and therefore possibly incorrect, alignments hypothesized by GIZA++. For a given French word, the count of the most frequent English alternative is divided by an alignment cutoff parameter k, and any English alternatives with counts less than the resulting value are removed from the list of alignments. In the example of Figure 1, the list of English translations for the French word *paru* would be pruned as shown in Figure 2 for different values of k.

During system development, the best results were found with a setting of $k = 2.5$. In the example of Figure 1, this preserves the generally-accepted translations of "appeared" and "seemed" for *paru*, but prunes out the secondary meaning "published," which is also a correct translation.

2.3 Bilingual Lexicon

A large word translation lexicon was then automatically produced using the filtered set of alignments. First, both the French and the English training corpora were tagged with the part-of-speech tagger TreeTagger

[1] `http://www.statmt.org/europarl/`

[2] A description of the translation task can be found at `http://www.statmt.org/wmt07/shared-task.html`.

French	English	Count
paru	appeared	27
paru	seemed	27
paru	found	10
paru	published	9
paru	felt	7
paru	struck	5
paru	thought	3
paru	was	3
paru	find	2
paru	seem	2
paru	already	1
paru	call	1
paru	deemed	1
paru	greater	1
paru	impression	1
paru	like	1
paru	occasion	1
paru	press	1
paru	release	1
paru	saw	1

Figure 1: Extracted alignments, and their frequency counts, for the French word *paru*.

Cutoff	Min Count	Filtered Alternatives
$k = 2.5$	$27/2.5 = 10.8$	appeared, seemed
$k = 5$	$27/5 = 5.4$	appeared, seemed, found, published, felt
$k = 10$	$27/10 = 2.7$	appeared, seemed, found, published, felt, struck, thought, was

Figure 2: Filtered alternatives for the French word *paru* given various alignment cutoffs.

(Schmid, 1994; Schmid, 1995), another freely-available online resource[3] that has been used for a variety of European languages. TreeTagger's part-of-speech sets are different across languages, but these differences can actually be useful in the lexicon creation process. French nouns, for example, all receive tags of NOM regardless of whether they are singular or plural; English nouns, on the other hand, will be marked as NN if singular and NNP if plural. Therefore, if the word alignments are assumed to be correct, information about the number of a French noun can be propagated from the English translation aligned to it in the corpus.

Given as input the part-of-speech tagged corpora and the filtered set of alignments, a series of lexicon-building scripts (one per system part of speech) produces lexical entries in the AVENUE transfer format. An entry is created from a word alignment if and only if the part-of-speech tags found in the corpus for both the French and English words can be collapsed to the same system-level part of speech. The output entry also contains any lexical features that can be induced from the French or English tags; an overview of these features is given in Figure 3.

English POS	**French POS**	**System POS**	**Features**
JJ*	ADJ	ADJ	*none*
RB*, WRB	ADV	ADV	*none*
IN*, TO*, RP*	PRP	P	*none*
NN, stem unknown	NOM, stem unknown	NAME	*none*
NN	NOM	N	num = sg
NNS	NOM	N	num = pl
V*	VER:infi	V	*none*
V*	VER:pres, VER:impi, VER:subp	V	tense = pres
V*	VER:ppre	V	tense = pres, aspect = imperf
V*	VER:simp, VER:pper	V	tense = past
V*	VER:impa	V	tense = past, aspect = imperf
V*	VER:subi	V	tense = past, aspect = imperf
V*	VER:futu	V	tense = future
V*	VER:cond	V	tense = cond
VB*, VH*	VER*	V	aux = +

Figure 3: Part-of-speech collapsing and lexical feature induction as carried out by the system's lexicon generation scripts.

The automatically-generated lexicon was supplemented with a comparatively small number of manually-written entries. These mostly cover closed-class categories such as determiners (DET), conjunctions (CONJ), negation words (NE and NEG), relativizers (REL), pronouns (PRO), and French preposition-plus-determiner combinations such as *aux* and *du*. Words in these categories are limited in number and carry a much richer syntactic feature structure than open-class words, so it was deemed advantageous to create more completely-specified entries by hand for them. The high frequency of function words in most input also provided motivation for writing entries for those words by hand in order to ensure that their English translations are correct. The manual lexicon also includes a small number of entries for specific sets of open-class words, such as the days of the week (as nouns) and the cardinal numbers from one to nine (as adjectives). Though these words should in theory be covered by the automatically-generated lexicon, they also are common enough in Europarl input that it was thought useful to have perfectly correct manual entries for them.

Figure 4 shows the final size of the word lexicon.

2.4 Noun Phrase Translation

As mentioned previously, an additional goal of this project was to take advantage of the consistency of noun phrases (NPs) across languages and improve overall performance by producing better NP translations.

[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

POS	Automatic # Entries	Manual # Entries
ADJ	13,697	10
ADV	1140	
CONJ		4
DET		43
N	45,878	7
NAME	18,669	
NE		2
NEG		6
P	90	10
PRO		49
REL		27
V	32,937	12
Total	112,411	170

Figure 4: Size of the word lexicon by part of speech.

Development efforts in this category are based on work previously carried out by Sanjika Hewavitharana, a member of the Carnegie Mellon statistical machine translation group, as part of a laboratory exercise.

For the current project, Hewavitharana provided a list of parallel French–English NPs extracted from 688,000 sentences of the Europarl corpus (Release 2) that had been parsed in English by Chris Callison-Burch. First, the English and French parallel texts were word aligned with GIZA++. Then, minimal NPs — defined as those that do have have smaller NPs nested within them — were found in the parsed English sentences, and their bounds were projected into the parallel French sentences based on the GIZA++ word alignments. Finally, the paired NPs were extracted and returned.

As in the case of the word-level alignments, the NP alignment data was also found to be noisy, so additional filtering steps were applied. Extracted NPs were thrown out if they consisted of single words, were wholly digits, contained punctuation, or if the French text consisted merely of "stranded" words such as variants of "a" and "of the." Phrases satisfying all these criteria were further filtered based on frequency count in the corpus and length ratio.

The filtered NP list was then added to the system as a phrasal lexicon without modifying the original word-level lexicon, thus allowing the creation of additional translation possibilities in the transfer lattice. The French NP *une motion de procédure*, for example, can still be translated word-by-word to produce "a point of procedure," but since the entire NP is also an entry in the phrasal lexicon, the (improved) English output "a procedural motion" is also possible.

The final NP lexicon built as described above contains 18,633 entries.

2.5 Transfer Grammar

The system's transfer grammar consists of 48 manually-written rules for combining lexical items and constituents into larger constituents, subject to a series of feature unification constraints. Many of the rules, specifically those building from adjectives and nouns, are based on the theory of X-bar syntax as explained, for example, by Radford (1988). Verb rules are built around the process of begining with a main verb (marked as V), possibly combining with auxillaries and negation words to form a verb cluster (marked VERB), and finally picking up a series of NP or PP arguments to form a verb phrase (VP).

Many grammar rules capture structural divergences between French and English, such as reordering of pronounal direct and indirect objects or post-nominal adjectives, but a number of rules also exist to provide basic coverage of syntactic structures. Sentence-level rules for imperatives (S → VP) or relative clauses (S → S REL S), for example, are included even though no reordering or feature unification is carried out within

them. In certain cases, these rules are necessary to create consitutents that will be used as input for more interesting higher-level rules. A series of consecutive proper names, for example, can be parsed into a name phrase (NAMEP), and a name phrase can be promoted to a noun phrase, which can then participate in sentence- or verb-phrase-level rules for subjects and objects.

Negation, which in French consists of two words (*ne ... pas* or *ne ... guère*, for example) surrounding an auxillary or main verb, is handled by two grammar rules that look for the initial *ne*, the correct type of verb, and an appropriate negation word (such as *pas* or *guère*). The English translation deletes *ne* and replaces the negation word with its equivalent (such as "not" or "hardly").

3 Examples

Further characteristics of the transfer grammar can be highlighted by examining a few parsed examples. A synchronous parse of a simple French N-bar and its English translation is given in Figure 5.

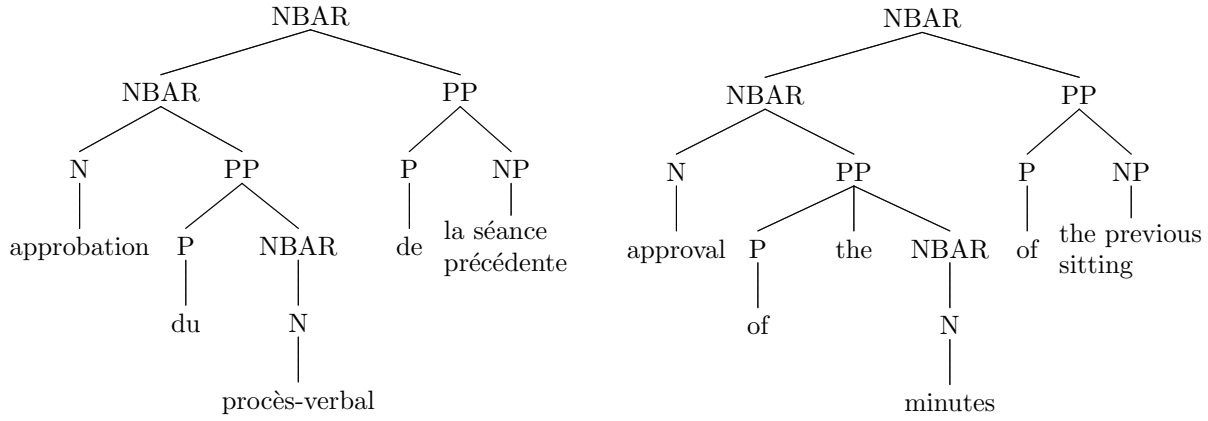

Figure 5: Synchronous parse and English translation generated for the French fragment *approbation du procès-verbal de la séance précédente*.

Of particular linguistic note in the example of Figure 5 is the handling of the structurally dissimilar prepositional phrases *du procès-verbal* and "of the minutes." While many French PPs have the familiar P NP structure as in English, there are also four preposition-plus-determiner combination words (*au*, *aux*, *du*, and *des*) that break the separation between the P and NP constituents. The French preposition *à* or *de* and the masculine determiner *le* or the plural determiner *les* from the following noun phrase combine to form a single token. In these cases, the structure of the French PP is more accurately expressed as PDET NBAR, where PDET is a preposition–determiner compound and NBAR is a noun phrase missing a determiner.

Synchronously generating this type of PP in the current system involves both the manual lexicon and the grammar. Lexical entries for *au* and *aux* are provided with the English translations "to," "in," or "at," and lexical entries for *du* and *des* have the English translations "of" or "from." All of these preposition entries are marked with a feature, `(detr +)`, on the French side indicating that their forms already include a determiner. In the grammar, a PP rule is added whose French right-hand side is `P NBAR` and whose English right-hand side is `P ‘‘the’’ NBAR`. Within the rule's body, a unification constraint specifies that the rule may only apply when the French-side P is marked as `(detr +)`. This correctly represents the input structure in French and produces the correct output text in English.

Figure 6 shows a more complicated sentence fragment.

A key step of the translation in the Figure 6 example is carried out at the VP level, where the French pronounal direct object *l'* ("it") and indirect object *vous* ("to you") are reordered to their correct positions in English. This type of reordering is only necessary — and permissible — with pronoun objects; in a fully-specified French sentence, such as *j'ai dit la réponse au professeur*, the order of the verb arguments remains

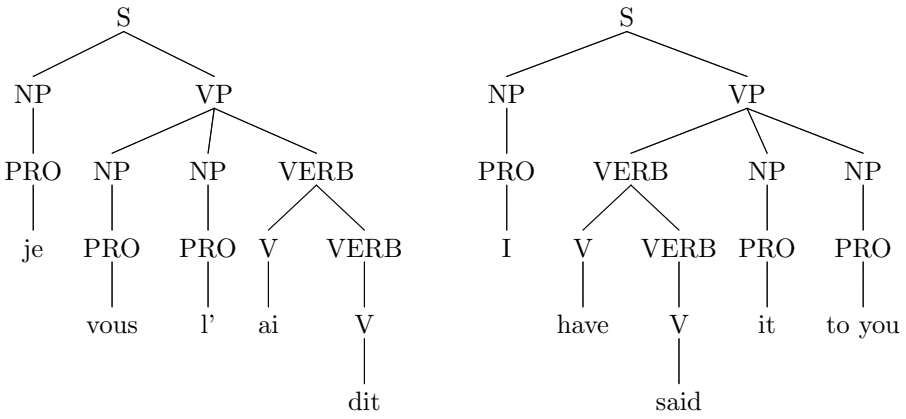

Figure 6: Synchronous parse and English translation generated for the French fragment *je vous l'ai dit*.

the same in the English equivalent ("I told the answer to the professor"). Verb-phrase-level rules that permit reordering thus include feature constraints to ensure that the NP objects are marked as pronouns and that the pronouns have the correct grammatical case. (Case is marked as a feature in the manually-generated lexical entries for pronouns.)

4 Results

In accordance with common practice, the Europarl transcripts covering October through December 2000 were reserved as development and testing data. From this, a specific development test set of 1073 sentences was created from the document for October 2, 2000. The first 30 sentences of the document were used as an incremental development set so that system progress and linguistic coverage could be quickly evaluated against a small sample of data.

Figure 7 shows final system results on the 1073-sentence development test broken down by system configuration. Scores are reported for the METEOR (Banerjee and Lavie, 2005) and BLEU (Papineri et al., 2002) automatic metrics. METEOR results were obtained with the exact match, Porter stemmer, and WordNet synonymy modules; BLEU results are case-insensitive and are calcualted according to the corrected BLEU 1.04 script released by IBM.

System Components	**METEOR**	**BLEU**
Word lexicon only	0.4289	0.1214
Word lexicon + grammar	0.4622	0.1540
Word lexicon + grammar + NP lexicon	0.4727	0.1613

Figure 7: Comparison of METEOR and BLEU scores on Europarl development data for various system configurations.

To provide an idea of "competitiveness," the system was also compared against the 10 translation engines that participated in the shared task of the 2006 ACL Workshop on Statistical Machine Translation (Koehn and Monz, 2006). Performance was evaluated on both the in-domain (2000 sentences from the Europarl corpus) and out-of-domain (1064 news commentary sentences) test sets. A summary of the results is given in Figure 8.

As a rule-based engine, the system created for this project shows less of a drop in BLEU score when moving from in-domain to out-of-domain data than do most statistical translators. The nine statistical

System	In-Domain	Out-of-Domain
Best 2006 System	0.3081	0.2195
Average 2006 System	0.2885	0.2057
Worst 2006 System	0.2144	0.1555
Current System	0.1770	0.1402

Figure 8: Comparison of BLEU scores between this system and systems submitted to the 2006 ACL shared translation task.

engines in the 2006 evaluation lost an average of 0.0898 BLEU when translating news commentary data as compared to Europarl data, while the single rule-based system fell 0.0202. The drop of 0.0368 shown by the current system is between the two ranges, but much closer to that of the rule-based system, as expected.

5 Analysis

The relatively stable performance on both in- and out-of-domain data indicates that the system is providing some payoff as a viable translator. However, the low range of the scores presented in the previous section shows that various aspects of the current implementation could be improved through additional development work or the application of new techniques. In the following sections, some of these aspects are highlighted and possible solutions are explored.

5.1 Word Alignment Cardinality

As mentioned previously, using the intersection of the GIZA++ French-to-English and English-to-French word alignments to build the system lexicon has the side effect that all lexical entries are constrained to map exactly one French word to exactly one English word. This can especially be a problem in capturing verb tense information. For future, conditional, imperfect, or infinitive forms, single-word French verbs (e.g. *prendra*, *aurait*, *parlais*, or *dire*) often must be expressed in English as two words ("will take," "would have," "was speaking," or "to tell"). On the other hand, simple past-tense verbs in English (e.g. "spoke") require two words in French (*a parlé*).

Since the input is in French, the second case can be handled easily in the grammar with a rule that allows an auxillary to be dropped when translating to English. Thus, a French verb cluster such as *ont bombardé des cibles*, which normally would produce "have bombed targets" in English, can also be translated to "bombed targets" as well.

The first case, however, is a more pervasive problem, since the one-word-to-one-word alignment constraint prevents multi-word English translations. In the word lexicon, the 122 first-person singular conditional verbs (ending in *-erais*) in French all have English translations consisting of only a main verb, so the necessary auxillary "would" is never produced. Of the 1009 entries for third-person singluar future-tense verbs (ending in *-era*), 945 of them have English translations that are only main verbs (lacking "will"); the remaining 64 translate as only "will" and are lacking the main verb. As a result, translation output sometimes appears to be missing words in English when the verb translations are not correct.

An attempt was first made to correct the problem by writing additional verb rules in the grammar. For a lexical V marked with `(tense future)`, a rule exists that leaves the French side as is but inserts the word "will" before the verb on the English side. Similar rules were written for verbs marked `(tense cond)`, adding "would" in English, and verbs marked `(tense *UNDEFINED*)`, adding "to" in English. Unfortunately, parse traces for the highest-scoring 100 translations for each sentence of the 1073-sentence development set show no occurrences of these rules being used.

A second solution was built from the original French-to-English GIZA++ alignments, which are one-to-many and thus can capture an alignment between a single-word French verb and a multi-word English equivalent. The French-to-English alignments were filted as described previously with an alignment cutoff

of $k = 2.5$, and then lists of French–English alignments where the English side begins with "to," "will," or "would" were extracted. The "to" list contained only 14 alignments, all of which were verified incorrect by hand, so they were discarded. The 2605 "will" alignments and 1378 "would" alignments, on the other hand, were converted to lexical entries with the part of speech V and added to the word-level lexicon. The additional entries improve system performance to 0.4819 METEOR and 0.1650 BLEU, a gain of 0.0092 METEOR and 0.0037 BLEU over the full-system results given in Figure 7.

5.2 Parse Fragmentation

Parse chunk information provided by the AVENUE transfer framework shows that output sentences are in general broken into a large number of sub-sentence units instead of a single unified parse. A large contributing factor to this effect is the fact that punctuation marks were broken off of adjacent words in the training data, thereby making them "words" of their own. However, there are no parts of speech assigned to them and no grammar rules that cover them. This means that any punctuation marks encountered in new input data are treated as unknown words and that a sentence will have at least as many disjoint parse chunks as the number of punctuation marks it contains.

To see how much of the parse fragmentation is being caused by punctuation marks, all punctuation was removed from the development set, and the punctuation-free input was translated again. The distribution of the number of parse chunks per sentence, both with and without punctuation, is shown in Figure 9.

# Chunks	**Punct.**	**No Punct.**
1–5	98	159
6–10	223	273
11–15	239	239
16–20	185	171
21–25	121	109
26–30	78	51
31–35	53	24
36–40	26	25
41+	50	22
Total chunks	18,774	15,497
Average	17.5	14.4

Figure 9: Distribution of number of parse chunks across sentences in the development set, both with and without punctuation included in the input.

Although the average number of parse chunks is reduced with punctuation removed from the input, most sentences are far from being represented as a single chunk or even a small number of chunks. Therefore, the best way to reduce fragmentation is probably to further extend the transfer grammar to capture a larger variety of source-language input structures. Parsing adverbs as components of verb phrases, for example, or restricting the feature constraints of the S $\rightarrow$ VP imperative rule to keep it from overapplying, would help reduce overall fragmentation. Fragment information provided by the AVENUE transfer framework can also help pinpoint the most necessary revisions or additions.

5.3 Lexical Coverage

The degree of lexical coverage of the system can be approximated from parse chunk information provided by the AVENUE transfer framework. The final translations for the 1073-sentence development set are made up of 29,281 lexicon-level tokens, of which 4181 (14.3 percent) were not covered by either the word-level or NP lexicon and were assigned by default the part of speech UNK by the framework. This figure includes, however, a large number of entries for punctuation marks — which were not modeled by the lexicon or

grammar and are therefore guaranteed to be unknown — and for tokens consisting entirely of digits and hyphens. (The latter case is common for the official designations of European Parliament bills and reports.) Ignoring these unknown tokens, a more representative out-of-vocabulary (OOV) statistic is 865 tokens, or 2.95 percent.

The difference between UNK and OOV tokens is striking: although there are only seven sentences in the development set that have no UNKs at all, there are 581 that do not contain OOVs.

System performance improves slightly when the development set is trimmed to only these 581 non-OOV sentences, with scores reaching 0.4816 on METEOR and 0.1644 on BLEU. This represents a METEOR improvement of 0.0089 and a BLEU improvement of 0.0031 over the full development set. The small size of the improvement, coupled with the relatively stable performance of the system on out-of-domain data as noted previously, suggests that translation performance is overall not much hindered by out-of-vocabulary words.

However, it was also noticed that the system can be incomplete or incorrect in its lexical coverage of French verb forms, leaving valid conjugations that were not seen in training data with no English equivalents even though other forms of the same verb do appear in the lexicon. Furthermore, some forms that do exist in the lexicon may have their English translations in an incorrect English conjugation, so that resulting target sentences are ungrammatical. (An example of this second situation can be seen in the system's translation of *je souhaite qu'elle multiplie ses efforts* as "I hope that it increasing its efforts," where the third-person singular verb *multiplie* has been incorrectly rendered as "increasing" rather than "increases.") A probable solution to this problem would be to implement morphological analysis on the training data to reduce all verbs to their root forms before the word alignment stage. A French morphological analyzer and an English morphological generator could then be plugged into the system at run time to produce more systematically correct English output. This approach was not implemented in the current system because of time constraints.

5.4 Noun Phrase Representation

Entries from the NP lexicon made up 2844 (9.71 percent) of the 29,281 lexicon-level tokens in the translation of the development set. When this figure is computed over the 100 best translations for each development sentence, the percentage falls to 9.50 percent. This indicates that the NP entries are in general slightly improving translation quality, at least in terms of the scores assigned within the translation framework, since their appearances are more concentrated towards the top of the n-best lists. The conclusion is supported by the boost in translation score provided by the NP lexicon, as shown in Figure 7; a future oracle experiment, in which appearance in a translation of items from the NP lexicon can be checked for correlation with the score of that translation according to METEOR or BLEU, could provide additional information.

6 Conclusions

This project succeeded in its primary goal of demonstrating that statistical techniques can be combined with a rule-based translation framework to rapidly produce a functioning MT system for a new language pair. The current system was planned, developed, and incrementally improved over a period of approximately eight weeks. Though its results fall behind those obtained by more state-of-the-art statistical and rule-based systems, it does represent a good start towards the construction of a competitive translator.

The importance of NPs as a translation unit has also been confirmed, as the addition of the NP lexicon as described provided a noticible boost in translation scores. Further investigation into obtaining high-quality parallel NPs, filtering them appropriately if necessary, and integrating them into the translation lexicon could additionally improve overall system performance.

With regard to NP translation as a subtask, a second stage of experiments is currently underway with data provided by Erik Petersen, the maintainer of the AVENUE transfer framework. The new parallel NP data comes from a crawl of the Wikipedia online encyclopedia system[4], in which the title of a French

[4] `http://www.wikipedia.org`

Wikipedia page and the title of its equivalent English page are extracted as one piece of parallel data.

Finally, the system analysis presented in this report suggests some fruitful directions for future work, specifically the development of a more thorough word alignment and lexicon-building process that is not restricted to one-word-to-one-word alignments, further enrichment of the transfer grammar, and the introduction of morphological analysis and generation. In addition, a refinement of both the French and English training data tokenization schemes (particulary with regard to hyphenated words) and experiments with different part-of-speech taggers and induction of lexical features represent further avenues for system development and improvement.

References

Banerjee, Satanjeev and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Koehn, Philipp. 2003. *Noun phrase translation*. Ph.D. thesis, University of Southern California.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. MT Summit.

Koehn, Philipp and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. *Proceedings of ACL 2003*, pages 311–318.

Koehn, Philipp and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121.

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

Papineri, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Peterson, Erik. 2002. Adapting a transfer engine for rapid development machine translation. Master's research paper, Georgetown University.

Radford, Andrew. 1988. *Transformational grammar: A first course*. Cambridge University Press.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT Workshop*.