Machine Translation for Languages Lacking Bitext via Multilingual Gloss Transduction

Brock Pytlik
Department of Computer Science
Johns Hopkins University
bep@cs.jhu.edu

David Yarowsky
Department of Computer Science
Johns Hopkins University
yarowsky@cs.jhu.edu

Abstract

We propose and evaluate a new paradigm for machine translation of low resource languages via the learned surface transduction and paraphrase of multilingual glosses.

1 Introduction

The dependence of traditional statistical machine translation on large training bitext presents a significant problem for low-resource languages where such bitexts are not typically available and are expensive to produce.

We propose a novel two-step approach to develop machine translation capabilities for a given source and target language, without *any* training bitext for that language pair. Our approach relies instead on a translation lexicon along with bitext between the target language and one or more supplementary source languages with characteristics similar to those of the main source language. We explain the approach by an example (outlined in Figure 1) in which Spanish is to be translated into English, using only French-English and Italian-English bitext. The approach relies on glossers, which we defined as simple translation systems informed only by translation lexicons rather than bitext. Glossers produce largely word-by-word translations, in approximately source-language word order. The first step is to apply the French and Italian glossers, yielding bitexts between English and an intermediate language, which we call *glossese*. The glossese is largely English in vocabulary but French and Italian in word order and idiom. The second step is to use these bitexts to train a glossese-to-English phrase-based machine translation system. Finally, to translate from Spanish to English, the Spanish glosser is applied to produce glossese, and the glossese-to-English system is applied to produce English.

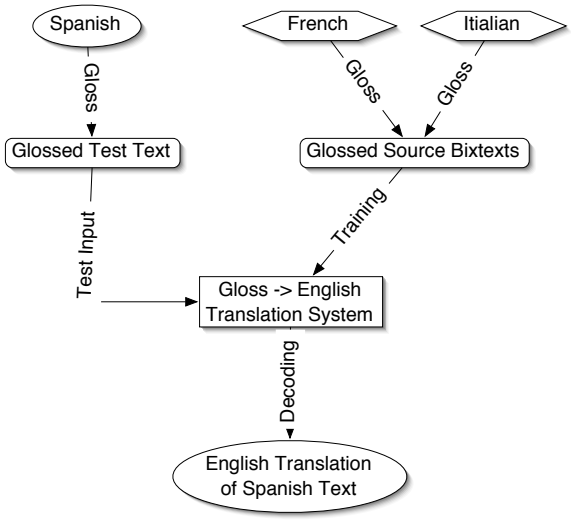

Figure 1: System Architecture

The glossing step is critical because the bitexts for various languages must be transformed into a common vocabulary and word order as much as possible. If the glosses produced from different languages are similar in these respects, then the phrases extracted by the machine translation system will be useful for other languages with similar word orders. For example, if the phrase *the house white* (possible as a gloss of *la casa blanca*) has been seen frequently in

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 156-165,
Cambridge, August 2006.

the glossed bitexts, there is a good chance the machine translation system has learned that *the house white* should be translated as *the white house*. If the test language fragment is glossed as *the house white*, then the system will work. If instead the test language is glossed as *white the house* or *the apartment white* then the machine translation system will be unlikely to correctly reorder the words.

We will first propose several simple methods of glossing which ignore surrounding context. We will also propose two glossing methods which take into account neighboring context as well as monolingual information for the target language. These systems focus on improving the choice of words during glossing to increase the vocabulary overlap among languages. Finally, we will propose a glossing method which can locally reorder words while glossing.

Again, the key to this work is that it does not require a bitext between the test language and the target language.

2 Related Work

Dirix et al. (2005) propose a system for machine translation without bitext. The system, Metis-II, assumes a great deal more resources than our system, including a chunker or parser and a part-of-speech tagger in the source language. Like our system, they assume the existence of a bilingual dictionary. They also use context, although with more intensive linguistic preprocessing, to choose a translation. Badia et al. (2005) also propose a system for machine translation without bitext. They also require a part-of-speech tagger and lemmatizer like Dirix et al. (2005). They use an n−gram language model to choose between various options but also rely on the POS tagger to guide their choice. Lavie et al. (2003) take a different approach for low resource languages. They use small amounts of word-aligned data and elicit information from a native speaker. By learning the structure of one language, they are able to transfer knowledge to a resource poor language, which provides a better translation.

3 Glossers

The dashed box in Figure 2 indicates where glossers are located in our system. Glossers take a text in one language and substitute (using a dictionary or other approaches) words in another language.

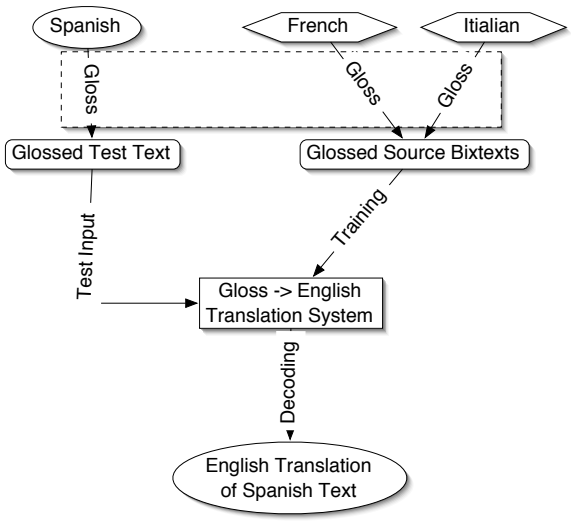

Figure 2: System Architecture Illustrating where Glossing Occurs

Glosses are differentiated by (at least) two axes: the morphological analyses performed and the method a glosser uses to choose among candidate glosses. In our glossers, we set aside the issue of morphological analysis and have focused instead on how to choose among several options to gloss a particular word. All of our glossers use the bilingual dictionary to establish the candidate glosses of a word.

Let T be the target language we wish to translate into. Let $\mathcal{L}$ be set of languages for which bitext (b_L) with T is available. For each $L \in \mathcal{L}$ we assume that a set of dictionaries exists between T and L (D_L). A glossing function ($g(w_L, L) \rightarrow w_G$) maps a word in L (w_L) to a word in glossese (w_G). Let $\gamma(w)$ be the set of words given as translations for $w \in L$ in D_L and $\gamma_d(w)$ be the words given as translations for a particular dictionary d in D_L.

3.1 Highest Frequency (HC)

The simplest approach for choosing among the gloss candidates is simply to choose the one with the highest frequency in target language monolingual text. Let $c(w)$ be the count of w in m_T where m_T is

monolingual text for T.

$$g_{\text{HC}}(w, L) = \arg\max_{x \in \gamma(w)} c(x) \quad (1)$$

This approach tends to conflate related concepts but instead focuses on producing a glossese with a relatively small vocabulary. For example, the following Spanish words are all glossed as *table* under this model:*abatible, aplazar, archivar, baremo, carpeta, liga, meseta, posponer, tablero, índice*.

3.2 Dictionary Frequency (DC)

If multiple dictionaries for T and L exist, another approach is possible. The glosser selects the candidate which is a candidate in the largest number of dictionaries.

$$g_{\text{DC}}(w, L) = \arg\max_{x \in \gamma(w)} \sum_{d \in D_L} I(x \in \gamma_d(w)) \quad (2)$$

This technique tends to select more specific words than HC. This may result in lower overlap between glosses of different languages. The value of the overlap may increase since the more specific gloss is being chosen more frequently. This model glosses the following words as *table*: *abatible, clasificación, cuadro, disponer, liga, mesa, posponer, tabla, índice* .

3.3 Language Frequency (LF)

A similar approach is to count the number of source languages where a word w appeared in a dictionary.

$$g_{\text{LF}}(w, L) = \arg\max_{x \in \gamma(w)} \sum_{L' \in \mathcal{L}} I(x \in D_L) \quad (3)$$

The approach focuses primarily on increasing the size of the overlap of glosses from different source languages to facilitate cross-lingual leveraging. The words which are glossed as *table* are: *abatible, aplazar, clasificación, cuadro, disponer, liga, mesa, meseta, posponer, índice*. This is essentially a mix of the words chosen by HC and DC.

3.4 Comparison of DC to HC and LF

Table 1 contains examples of which candidate each glosser chose for several different Spanish words. Comparing HC and DC, it appears that when they disagree, HC chooses a more general word while DC chooses a more specific word. It is less clear what pattern exists between LF and DC.

Spanish Word	Basic Glossing Approach		
	HC	DC	LF
abadejo	cod	swordfish	pout
abandonado	abandoned	vacant	neglected
cartilla	book	primer	book
crio	child	brat	child
domiciliario	house	domiciliary	house
gallinero	pen	henhouse	pen
gasolina	gas	petrol	juice
plumaje	feathers	plumage	plumage
porcino	pig	porcine	bump

Table 1: Examples of Different Candidates Selected by Glossers

3.5 Disambiguation Based Approaches

A more complex model is to use the monolingual information provided by target language monolingual text (m_T) to disambiguate during glossing. For a given word w_i in m_T take the first non-function word to the left or right of it as the context (w_p, w_n). Function words are skipped to allow the glosser to focus on relations among content words. For all combinations of the glossing candidates for w_i and w_p or w_n, in either order, count the number of times each glossed bigram occurs in m_T. Choose the candidate for w_i which occurs the most times over all the bigram combinations. For example, suppose a Spanish sentence contains the sequence ... *gobiernos de los estados miembros* ... and the system needs to determine how to gloss *estados*. *Gobiernos* and *miembros* are the relevant context since the function words *de* and *los* are ignored. For the purposes of illustration, two possible dictionary-based glosses of *estados* are *states* and *report* while the possible glosses of *gobiernos* include *controls* and *government*. *Miembros* has possible glosses *member* and *fellows*. Table 2 shows the frequency of the candidate word combinations in English (in both orders and skipping intermediate function words). *States* is selected as the preferred gloss for *estados* because it has been seen with more possible glosses of *gobiernos* and *miembros* than *report*.

This method disregards word order for both languages.

3.5.1 Word Based Word Sense Disambiguation (WWsd)

WWsd is the formal implementation of the ideas presented above. Let $\phi_T(w, w')$ be the count of the

Disambiguating *estados* in *gobiernos de los estados miembros*

Source Words	Candidates for glossing	
gobiernos	controls	government
estados	**states**	*reports*
miembros	member	fellows

Left Spanish Context				Right Spanish Context			
controls **states**	0	**states** member	218	controls *reports*	0	*reports* member	88
states controls	1	member **states**	124418	*reports* controls	1	member *reports*	5
government **states**	23	**states** fellows	0	government *reports*	1	*reports* fellows	0
states government	55	fellows **states**	0	*reports* government	3	fellows *reports*	0
Total **states**: 124715				Total *reports*: 98			

Disambiguating *comunidades* in *general de las comunidades europeas*

Source Words	Candidates for glossing	
general	prevailing	general
comunidades	**communities**	*common*
europeas	european	

Left Spanish Context				Right Spanish Context			
prevailing **communities**	0			prevailing *common*	0		
communities prevailing	0	**communities** european	229	*common* prevailing	0	*common* european	512
general **communities**	2	european **communities**	152480	general *common*	9	european *common*	32
communities general	61			*common* general	1		
Total **communities**: 152772				Total *common*: 554			

Disambiguating *conformidad* in *programa de conformidad con el procedimiento*

Source Words	Candidates for glossing	
programa	plan	programme
conformidad	**accordance**	*conformity*
procedimiento	method	procedure

Left Spanish Context				Right Spanish Context			
plan **accordance**	39	**accordance** method	141	plan *conformity*	3	*conformity* method	0
accordance plan	14	method **accordance**	11	*conformity* plan	14	method *conformity*	2
programme **accordance**	134	**accordance** procedure	6579	programme *conformity*	14	*conformity* procedure	29
accordance programme	18	procedure **accordance**	127	*conformity* programme	11	procedure *conformity*	11
Total **accordance**: 7063				Total *conformity*: 84			

Table 2: Examples of Monolingual Word Sense Disambiguation. For purposes of tractable illustration, the space of candidates is limited to 2.

bigram ww' in m_T. w_n and w_p are the next and previous words relative to w.

$$g_{\text{WWsd}}(w, w_n, w_p, L) = \arg\max_{g \in \gamma(w)}((\sum_{g_l \in \gamma(w_p)} \phi_T(g, g_l) + \phi_T(g_l, g)) + (\sum_{g_r \in \gamma(w_n)} \phi_T(g, g_r) + \phi_T(g_r, g))) \quad (4)$$

A variation on this method would take the maximum of the first row and the second row. This variation implies that for a given word pair, the gloss should either always or never be flipped in order.

One deficiency of WWsd is that since each word is glossed separately, the gloss of *comunidad* does not take into account the actual gloss of *abarcaba*, only its candidate glosses. The advantage of this is that it reduces a possible cascade of errors caused by a bad previous glossing choice. The disadvantage is that the bigram that is produced may not be the globally optimal bigram given both left and right context.

3.6 Phrase Reordering

All the glossers discussed up to now share a fundamental weakness: they produce glossed text in the same word order as the source language. Since languages of the world differ in word order in many ways (such as SVO vs SOV canonical word order, nouns preceding adjectives vs nouns following adjectives), our glosser should be able to reorder the glossed words to match the target language order.

Target	Source	Glossed
a blue rectangle and encircled	un rectángulo azul y rodeado	a rectangle blue and encompassed
limited implementation	aplicación limitada	application limited
is the first place	constituye el primer luga	form the first place

Table 3: Examples of Glossed Spanish Training Phrases Using WWsd

Fully reordering the glossed output would amount to having a full distortion model from source text to glossed text. Instead, we chose to only allow very local reordering of bigrams. The next glosser presented has this ability.

3.6.1 Phrase Based Word Sense Disambiguation (PWsd)

The final pure glossing technique conducts phrase based word sense disambiguation using monolingual text and a dictionary. The same distributions are measured as in WWsd except that only the left context is used. Rather than taking a sum or a max to choose a single word, a dictionary entry is created for the ordered pair of source words and the best ordered pair of target words. Continuing the previous example, instead of glossing *abarcaba* and *comunidad* separately as WWsd does, they would be jointly glossed as *included community*.

$$g_{\text{PWsd}}(w, w_n, L) = \underset{(g \in \gamma(w), g_n \in \gamma(w_n)) \cup (g \in \gamma(w_n), g_n \in \gamma(w))}{\arg\max} \phi_T(g, g_n) \quad (5)$$

Depending on how the function words which are skipped when choosing w_n are handled, it is possible for PWsd to delete function words during glossing. The consequences of this are addressed in Section 6.3.3.

3.7 Bitext Based Glossing (Bible)

We created one glosser which made use of a small amount of bitext. For several languages, we had a bitext of the Bible available. We trained Giza++ (Och and Ney, 2003) on each of these bitexts. Combining the bidirectional alignments using the grow-diag-final algorithm (Koehn et al., 2003), we created one-to-one mappings using the mostly likely target language word given the source language word. These mappings were treated as a dictionary and used to gloss text.

Abbreviation	Language
DE	German
ES	Spanish
FR	French
NL	Dutch
PT	Portuguese
SV	Swedish

Table 5: Abbreviations Used for Languages

4 Glossing Experiments

4.1 Data

Our experiments are performed on the Europarl-04 data. All data were tokenized using a tokenizer based on the one distributed with Egypt (Al-Onaizan et al., 1999) but extended to work on other languages. The evaluation data consists of every tenth sentence of the first half of the partitions after the 100th partition, resulting in between 600K and 700K words (20K sentences) per test set. Table 5 lists the languages used in the experiments and the abbreviations used for the languages.

4.2 Results

4.2.1 Baselines

To establish a baseline, we first treated the unaltered source language side of the test set as if it were the output of a glosser and evaluated it. This yields a BLEU score higher than 0 because of borrowed words and names. The top of Table 6 shows the results for three languages.

We established another baseline by using the bilingual dictionary as an (impoverished) bitext. A Giza++ (Och and Ney, 2003) system was trained on this bitext and Pharaoh (Koehn, 2004) was used to decode the test set[1]. Many languages have translations of the Bible available even when no other

[1]Section 6 gives more details on our machine translation system.

Target	Source	Glossed
alternative stage	fase alternativa	alternative stage
its representative the following document	su representante el documento siguiente	its representative the following document
the commission of the european communities	la comisin de las comunidades europeas	you commission of them europeans common
common organisation of the market	organizacin comn de el mercado	common organization of the market
annex to that regulation	anexo de dicho reglamento	annex of declaration regulation

Table 4: Examples of Glossed Spanish Training Phrases Using PWsd

Model	Source languages		
	ES	FR	SV
Baselines			
No Translation or Glossing	.094	.099	.094
Dictionary as bitext	.149	.117	.122
Bible as bitext	.099	.127	.079
Morph. Expanded HC	.143	NA	NA
Glossing Systems			
System HC	.135	.158	.134
System DC	.124	.154	.133
System WWsd	.147	.167	**.137**
System PWsd	**.161**	**.179**	**.137**
System Bible	.128	.165	.133
System LF	.113	.129	.125
System WWsdToBible	.154	.178	.136
Fully Supervised MT			
Pharaoh System (trained on 76M words)	.308	.309	.252

Table 6: BLEU Score Performance for Glossing

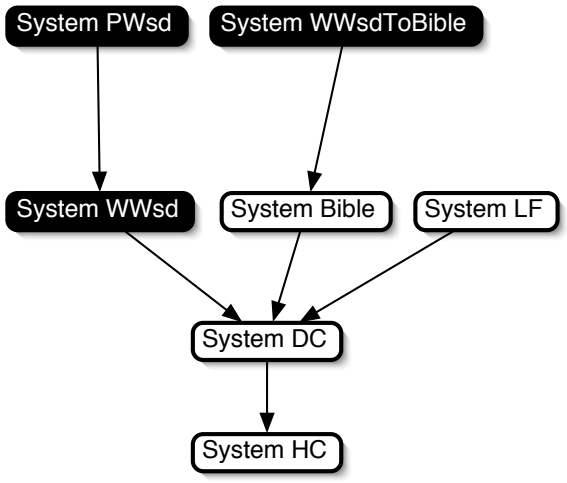

Figure 3: Depiction of Backoff Relationships

bitext exists. The same training and decoding procedure was used to build a system using the Bible bitext. Finally, the morphologically expanded Spanish glosser was evaluated in isolation, showing independent improvement due to better inflection handling. Also, a fully supervised Pharaoh system was trained on 76M words of bitext as a performance upper bound.

4.2.2 Systems

To combine glossing techniques, we used a cascade setup. For example, a glosser might first use the phrase level word sense disambiguation ($PWsd$) approach. If the word to be glossed cannot be disambiguated at the phrase level, then the bigram word sense disambiguation ($WWsd$) is used. If neither the left or right contexts have been seen with this word in training, then dictionary count (DC) and finally highest count (HC) glossers are used. Figure 3 shows the backoff relationship between various glossing systems. For example, the arrow from **System PWsd** to **System WWsd** means that **System PWsd** backs off to **System WWsd** when no bigram translation can be found.

A morphologically expanded dictionary was provided for Spanish. Those systems with black backgrounds used this expanded dictionary and should be compared against **Morphologically Expanded HC** instead of **System HC**.

4.2.3 Results and Analysis

The bottom half of Table 6 contains the results for these approaches. Systems are evaluated on their BLEU score using up to 4-grams. Table 6 indicates that three languages show sizable improvements relative to the baselines. Table 7 provides information about the dictionaries in question.

PWsd and WWsd are the most effective glossers tested. PWsd out-performed all baselines, including using the dictionary as a bitext. As expected, the results fall short of having large amounts of bitext in the desired language pair to train from. Nonetheless, the performance of the glossers gives hope that enough words are glossed correctly that the translation stage may be fruitful. The variance among the

Dictionaries	Entries	Words > 1 Entry	Words = 1 entry	Most Choices
ES	204,754	34,266	73,589	140
Expanded ES	1,117,142	192,508	301,341	185
FR	103,267	18,063	45,041	53
SV	116,784	22,196	58,012	49

Table 7: Dictionary Size and Fertility

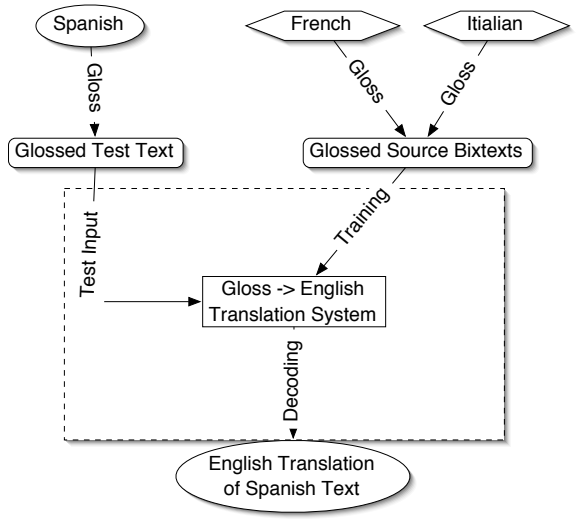

Figure 4: System Architecture Highlighting those Aspects Involved in Translation

glossers is also encouraging as it suggests that the type of glosser used matters.

5 Translation

Based on the success of glossing presented above, the next step was to use glossed bitext to train a machine translation system for translating common glossese into English. The goal was to improve the order of the English produced as well as possibly improving the word choice.

We propose a two-step system for machine translation when a bitext for the source language is not available. For each source language with bitext with the target language, the source language side is glossed. These glossed bitexts are combined into training for a single translation system. For our experiments, we used Giza++ to learn the alignments between glossed text and the target language, grow-diag-final (Koehn et al., 2003) to improve the phrases extracted, and Pharaoh to decode glossed text during testing. When translating a text for a language for which no bitext is available, the first step is to gloss the text using the same glosser that the training bitexts used. Pharaoh then decodes the glossed text into the target language. Figure 4 illustrates the system architecture. The parts enclosed in the dotted box are those described in this section. See table 8 for examples of the output of the decoding process.

We tried two different methods for combining the glossed text of the source languages. The first is to concatenate the glossed versions of the bitexts for the source languages (e.g. French and Italian). One potential advantage of this approach is that it mirrors the single language setting. Giza++ only needs to learn the alignment between glossed text and the target language. Learning the alignment over the aggregated languages may be more difficult than learning any single language alignment since various language-pair-specific tendencies will be competing in the distortion model as well as in the translation tables. This difficulty may improve performance because the alignments and the phrases extracted may be less likely to reflect the idiosyncrasies of any particular language. This may improve generalization to a new language. The practical disadvantage of this approach is that it produces large bitexts on which we were often unable to run Giza++ because of memory issues.

The second approach is to train Giza++ on the glossed text for each language separately, which avoids the previous practical disadvantage. After the phrase counts for each language are extracted, they are combined to produce a single phrase table. The possible advantage of this approach is that the alignments are more likely to be correct for each individual original language pair. The potential disadvantage is that the entries comprising the phrase table may be less likely to generalize to a new language. It

Original	Glossed	Translated	Reference
de el carbón y de el acero	of the coal and of the steel	of the coal and steel	coal and steel
el derecho de aduana preferencial	the law of customs preferencial	the customs duty preferencial	the preferential customs duty
una ' conductividad eléctrica en volumen '	a ' conductivity electrical to volume	a ' electrical conductivity to volume	a ' bulk electrical conductivity '

Table 8: Steps for Decoding a Test Sentence

is possible that some generalization will be achieved by the competition of target language phrases from different source languages which are translations of the same glossed phrase.

6 Translation Experiments

6.1 Data

As described in section 4.1, we used the Europarl-04 data. The training set for each language pair was created by using the first 100 of the provided partitions. We also removed all bitext sentence pairs whose lengths differed by more than a factor of two as well as all sentence pairs which were identical on both the source and target sides. For each language, approximately 76M words (1.7M sentence pairs) were extracted for use in cross-language training. The same evaluation sets that were used for evaluating glossers (see Section 4.1) were used to evaluate the translation systems.

6.2 Evaluation

We compare glossing performance by two measures. The first is the quality of the glossed text directly as translation candidates, evaluated against reference sentences. These results are presented in section 4.2. The second evaluation measure is the quality of the output of the Pharaoh machine translation system which uses this glossese as input. These results are presented in Tables 9 and 10. One hypothesis being tested is that using a phrase table trained on one source language to decode the glossing of another language will improve performance over the gloss alone. The second hypothesis is that using phrase tables from more than one source language will improve performance beyond a single language's phrase table.

6.3 Results and Analysis

Tables 9 and 10 are representative of the pattern seen across glossing techniques, source language combinations, and test languages. The last line of Table 9 shows the performance when the system is trained only on a glossed test language/English bitext. Using the translation system trained on one language after glossing uniformly degrades performance, given the word conflation and noise introduced into the original, with no source for potential countervailing benefit. However, combining results from multiple languages in combination may improve performance slightly as additional languages are added. For example, Table 9 shows NL performance after translation improving when FR is added to ES. However, the performance after translation never matches the performance of glossing alone. Thus the translation systems trained on glossese from other languages do not fully substitute for language-specific training data.

Sources	Test Languages		
	DE	NL	PT
Gloss Alone	**.128**	**.169**	**.092**
ES	.086	.111	.070
ES FR	.087	.141	.068
ES FR IT PT	.087	.120	*.300*
DE ES FR IT PT SV	*.209*	.126	*.298*
Target=Source	*.214*	*.246*	*.304*

Table 9: BLEU Scores after Translation when System HC is used to Gloss

Sources	Test Languages
	FR
Gloss Alone	**.1666**
ES	.1253
ES PT	.1259
ES SV	.1250
ES PT SV	.1250
	ES
Gloss Alone	**.1474**
FR	.1174
FR PT	.1209
FR SV	.1106
FR PT SV	.1151

Table 10: BLEU Scores after Translation when System WWsd is used to Gloss

Src	Tst	1grm	2grm	3grm	4grm	5grm
FR	ES	9292	28718	28998	16846	6979
FR	FR	18211	81974	110217	96693	65373

Table 11: Number of n-grams appearing at Least Once in Test

Src	Tst	1grm	2grm	3grm	4grm	5grm
FR	ES	1	3.09	3.12	1.81	.075
FR	FR	1	4.50	6.05	5.31	3.59

Table 12: Ratio of n-grams appearing at Least Once in Test

6.3.1 Phrase Table Details

One possible reason appears when the phrase table entries that are used during decoding are examined. Table 11 shows the number of distinct n-grams present in the French phrase table that are present in the test set. Not surprisingly, decoding cross-lingually results in fewer n-grams, across all levels. The drop of unigrams from 18211 to 9292 suggests a high OOV rate. It should be noted that while the test sets are not identical cross-lingually, they are very similar in both size and genre, which makes such comparisons meaningful. A more interesting trend is clear in Table 12. This table has been normalized by the number of unigrams found in the test set. In the cross-lingual situation, longer n-grams are disproportionally reduced. Averaged across all glossing, source, and test pairs for which data is available, this pattern holds. Table 13 demonstrates that the number of large n-grams seen in Tables 11 and 12 is actually a better than average result.

These numbers come from monotonic glossers. It is possible that a glosser which can locally reorder, like System PWsd, will not suffer as heavy a dropoff for longer n-grams, especially bigrams.

If unigrams do most of the work during decoding, this is not troubling, but if the longer n-grams do most of the (correct) work during decoding, then their failure to appear cross-lingually is of great concern. To test this, we built three simple decoders.

	1grm	2grm	3grm	4grm	5grm
Src≠Tst	1	2.74	1.68	0.63	0.24
Src=Tst	1	4.90	5.75	4.14	2.58

Table 13: N gram Ratio Averaged Across All Available Configurations

System	Src	Tst	Decoders		
			BG	BU	BF
WWsd	ES	FR	.083	.052	.167
WWsd	FR	ES	.086	.056	.147
WWsd	FR PT	ES	.085	.056	**
WWsd	FR PT SV	ES	.080	.055	**
WWsd	FR SV	ES	.080	.055	**

Table 14: Bigram Based Decoding

6.3.2 Bigram Based Decoding

The first decoder takes the phrase table which Pharaoh would use and removes everything but those entries with bigrams on the glossed side. The most likely translation for each gloss bigram is kept, creating a bigram dictionary. This dictionary is used in a greedy left-to-right replacement procedure. If a bigram on the gloss side matches a dictionary entry, the English side of the entry is used and the next word is skipped, having already been translated. We will call this decoder **BG**. The second decoder is identical to the first except that it also retains the unigram entries. During decoding, if a bigram for the current source bigram is not found, it uses a unigram entry for the left-most word of the source and then continues to the next word. We will call this decoder **BU**. The third decoder is identical to the first except that all entries containing function words on either the gloss or the English side are removed from consideration. This decoder is called **BF**.

Table 14 shows how BG, BU, and BF perform in various settings. The table shows that unigram entries in the phrase table hurt cross-lingual performance. This suggests that one approach for improving the performance of Pharaoh may be the removal of all unigram entries. A comparison of BG and BF also shows that a large gain is made by ignoring both unigrams and bigrams containing function words. That BF performance is close to the original gloss score is encouraging, because it is the best any decoding attempt using the phrase tables has done, but discouraging, because it suggests that BF is essentially doing no work and just passing the glossed text through. Nonetheless, these results suggest some possible approaches for improving the performance of decoding using Pharaoh in general.

Modification Type	Post Mod	Post Decode
No Change	1.00	.937
Remove `the, a, an`	.781	.821
Remove `of, in`	.835	.852
Remove `the, a, an, of, in`	.660	.720

Table 15: Pharaoh Performance Restoring Deleted Function Words (BLEU scores)

6.3.3 Restoring Function Words

A translation system trained on a PWsd gloss might need to regenerate deleted function words. We performed experiments to determine the difficulty of reinserting function words assuming an otherwise perfect translation. We took an English text and removed function words from it. We treated this modified text as the source text after it had been glossed and used the original text as the target language text. We trained our translation system and decoded using Pharaoh as described in Section 5. Table 15 shows the results, which are reassuring on several fronts. First, the reasonable performance on unmodified English text suggests that our configuration of Giza++ and Pharaoh is not badly broken. Further, in each of the three modified situations, using the translation systems improves the translation. This suggests that if System PWsd produces reasonable glossed non-function words, the machine translation system will be able to restore many of the missing function words.

7 Conclusion

We have proposed a novel approach to machine translation for languages without available bitext. The two-step process uses an existing dictionary and monolingual corpus information to transform input text into a common source language which we call glossese. We proposed several methods for glossing the source text which use the dictionary to constrain the possible gloss and use target language information to select a particular gloss. We explored some of the issues that arise when attempting to use phrase tables trained on the gloss of one language to decode the gloss of another language. We discovered that function words and unigrams were particularly poor for translating the gloss of other languages. We also discovered that longer n-grams are disproportionately unlikely to appear cross-lingually.

The initial results for the glossers are encouraging. The context sensitive glossers were able to properly disambiguate words and the PWsd glosser was sometimes able to correctly reorder the glossese. The analysis of the problems with running a phrase-based machine translation system cross-lingually suggests that some of these issues may be overcome by restricting the entries in a phrase table.

Acknowledgements

We are grateful to Elliot Drabek of Johns Hopkins University for his donation of his morphologically expanded Spanish dictionary.

References

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, I. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. 1999. Statistical Machine Translation - Final Report. JHU Workshop 1999. Technical Report, Johns Hopkins University. pages 1–42.

Toni Badia, Gemma Boleda, Maite Melero, and Antoni Oliver. 2005. An *n*-gram approach to exploiting a monolingual corpus for Machine Translation. In *Proceedings of the Second Workshop on Example-based Machine Translation*.

Peter Dirix, Ineke Schuurman, and Vincent Vandeghinste. 2005. METIS-II:Statistical Machine translation using monolingual corpora - System description. In *Workshop on Example-Based Machine Translation*.

Philip Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings NAACL/HLT 2003*, pages 48–54.

Philipp Koehn. 2004. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124.

Alon Lavie, Stephan Vogel, Lori S. Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjós, Rachel Reynolds, Jaime G. Carbonell, and Richard Cohen. 2003. Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario. *ACM Trans. Asian Lang. Inf. Process.*, 2(2):143–163.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.