Matador: A Large-Scale Spanish-English GHMT System

Nizar Habash
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20740
habash@umiacs.umd.edu

Abstract

This paper describes and evaluates Matador, an implemented large-scale Spanish-English MT system built in the Generation-Heavy Hybrid Machine Translation (GHMT) approach. An extensive evaluation shows that Matador has a higher degree of robustness and superior output quality, in terms of grammaticality and accuracy, when compared to a primarily statistical approach.

1 Introduction

This paper describes and evaluates Matador, an implemented large-scale Spanish-English MT system built in the Generation-Heavy Hybrid Machine Translation (GHMT) approach introduced in (Habash, 2002; Habash and Dorr, 2002). GHMT is an asymmetrical hybrid approach that addresses the issue of MT resource poverty in source-poor/target-rich language pairs by exploiting symbolic and statistical target-language (TL) resources. Expected source-language (SL) resources include a syntactic parser and a simple one-to-many translation dictionary. No transfer rules or complex interlingual representations are used. Rich TL symbolic resources such as word lexical semantics, categorial variations and subcategorization frames are used to overgenerate multiple structural variations from a TL-glossed syntactic dependency representation of SL sentences. This symbolic overgeneration accounts for possible translation divergences, cases where the underlying concept or "gist" of a sentence is distributed differently in two languages such as *to put butter* and *to butter* (Dorr, 1993). The overgeneration is constrained by multiple statistical TL models including surface n-grams and structural n-grams. The source-target asymmetry of systems developed in this approach makes them more easily retargetable to new source languages (provided a SL parser and translation dictionary).

An evaluation of Matador's translation quality is conducted using the IBM Bleu metric (Papineni et al., 2001) and comparing against three systems—simple gisting, primarily statistical (IBM Model 4) and purely symbolic (Systran)—over three corpora (UN, FBIS and Bible). The evaluation shows that although Matador scores lower than IBM Model 4 on the corpus where all language models were trained (UN), Matador has a higher degree of robustness and scores higher when tested on text with new genre (Bible). Additionally, the evaluation shows that Matador's output quality, in terms of grammaticality and accuracy, is superior to IBM Model 4.

The next section is an overview of Matador. This is followed by three sections corresponding to the three phases of GHMT: Analysis, Translation and Generation, respectively. And finally Section 6 presents an extensive evaluation of Matador.

2 Overview of Matador

Figure 1 describes the different components of Matador. There are three phases: Analysis, Translation and Generation. The last phase is marked as EXERGE — EXpansivE Rich Generation for English — a SL-independent generation module for English. These three phases are very similar to other paradigms of MT: Analysis-Transfer-Generation or Analysis-Interlingua-Generation (Dorr et al., 1999). However, these phases are not symmetric. Analysis relies only on the Spanish sentence parsing and is independent of English generation. The output of Analysis is a deep syntactic dependency that normalizes over syntactic phenomena such as passivization and morphological expressions of tense, number, etc. Translation converts the Spanish lexemes into bags of English lexemes. The dependency structure of the Spanish is maintained. The last phase, Generation, is where most of the work is done to manipulate the input lexically and structurally and produce English sequences. The generation resources are described next followed by an explanation of the generation submodules.

For example, the Spanish sentence *Maria puso la mantequilla en el pan (Mary put the butter on the bread)* is analyzed to produce a dependency tree, a representation describing the syntactic relations among the words in the sentence:

(1)
```
(puso :subj Maria
      :obj  (mantequilla :mod la)
      :mod (en :obj (pan :mod el)))
```

This dependency tree specifies that *Maria* is the subject of the verb *puso* and that *mantequilla* is the object. In the translation step, each of the Spanish words in the dependency tree are mapped into sets of English words:

(2)
```
((lay locate place put render set stand)
      :subj Maria
      :obj ((butter bilberry) :mod the)
      :mod ((on in into at) :obj ((bread loaf)
                                  :mod the)))
```

During generation, different variants of (2) are expansively created using lexical semantic information and other English-specific heavy resources. The following are only a few of these variants:

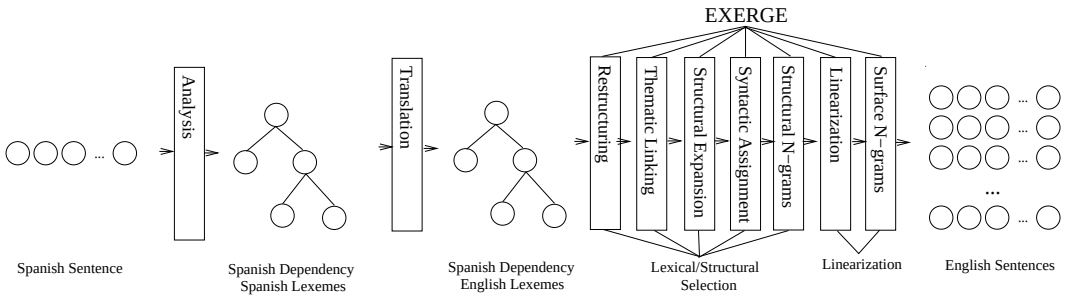

Figure 1: Matador: Spanish-English Generation-Heavy Machine Translation

(3)
```
(put  :subj Maria
      :obj ((butter bilberry) :mod the)
      :mod (on :obj ((bread loaf) :mod the)))
(lay  :subj Maria
      :obj  ((butter bilberry) :mod the)
      :mod (at :ob-
ject ((bread loaf) :mod the)))
(butter
      :subj Maria
      :obj ((bread loaf) :mod the))
(bread
      :subj Maria
      :obj ((butter bilberry) :mod the))
```

The first two examples show little difference in structure from the Spanish structure in (2), but the last two are much different. In the linearization step, the dependency trees in (3) are converted into a word lattice compressing multiple possible sentences:

(4)
```
(OR (SEQ Maria (OR puts put) the (OR butter
         bilberry) (OR on into) (OR bread loaf))
    (SEQ Maria (OR lays laid) the (OR butter
         bilberry) (OR at into) (OR bread loaf))
    (SEQ Maria (OR butters buttered) the
        (OR bread loaf))
    (SEQ Maria (OR breads breaded) the (OR butter
        bilberry)))
```

These different sequences are then ranked using a statistical language model. The overgenerated variants score higher than direct word translations, e.g., the top-ranked output in this example is *Maria buttered the bread*.

(5)
```
Maria buttered the bread          -47.0841
Maria butters the bread           -47.2994
Maria breaded the butter          -48.7334
Maria breads the butter           -48.835
Maria buttered the loaf           -51.3784
Maria butters the loaf            -51.5937
Maria put the butter on bread     -54.128
```

Matador uses some off-the-shelf components, namely the Connexor Spanish parser for analysis (Tapanainen and Jarvinen, 1997) and the Halogen Forest Ranker for surface N-gram ranking (in Exerge) (Langkilde, 2000). All other components were created or extracted as part of this research. An online-demo of Matador is available at http://clipdemos.umiacs.umd.edu/matador/.

3 Analysis

Spanish analysis in Matador utilizes the Connexor parser (Tapanainen and Jarvinen, 1997), a symbolically driven system based on Constraint Grammar (Karlsson, 1990). Connexor's output is a functional dependency that is somewhat incompatible with the input expected by Exerge. On the one hand, the functional dependencies for Spanish include thematic relations such as location and instrument. These relations are specified directly between a verb and its object regardless of the existence of a preposition. In this aspect, Connexor's output is deeper than what Exerge expects. On the other hand, some features closer to the surface form are kept such as complex verb chains signifying passivization. Other problems with parsing with Connexor include its "over-parsing" of complex untranslatable alphanumeric sequences, e.g., document references in the UN corpus such as *AA.33.1.C.ii/1991*, its failing to parse certain words altogether and its lack of handling of empty categories (i.e. `*trace*` and `*pro*`). Moreover, the set of parts of speech and relationship names are not consistent with what Exerge expects (including the handling of punctuation and conjunctions). For these reasons, the output of Connexor is further processed to make it Exerge-compatible.

The rest of this section focuses on four specific phenomena: auxiliary verb chains, reflexive clitic "se", depassivization and pro drop restoration. First, auxiliary verb chains are replaced with the features they specify such as perfect/progressive aspect or passive voice. For example, the auxiliary *estar* and the verb past-participle feature are replaced with the feature `(:voice passive)`. Features of auxiliaries are passed to the parent. A `*pro*` is added in subject position if there is pro-dropping. Although the auxiliary verb chain looks like English, it needs to be processed since the chain makes using subcategorization frames impossible. Moreover, this is problematic to the later step of structural expansion.

Second, the Spanish reflexive clitic "se" is used to indicate a variety of phenomena such as passivization, or emphasis (Garcia, 1975; Maldonado, 1988). In some cases, the meaning is changed in a pragmatic non-compositional manner. For example *acordar* is *to agree* but *acordarse* is *to remember*. Cases where meaning change occurs are indicated in the translation lexicon as separate entries from basic verbs. The reflexive clitic is treated in one of two ways depending on whether the infinitive form of the verb appears in the translation dictionary with the reflexive clitic attached.

- If the reflexive form of the verb appears, the clitic is attached to the verb.
- Otherwise, the clitic is deleted and the feature `(:voice passive)` is assigned to the main verb.

Third, in pro-drop languages such as Spanish, the subject pronoun of a verb can be dropped but is indicated in the morphology of the verb. To ensure that every verb has a subject, a place holder for the pro-dropped subject is added.

- If the verb has no subject and it is a child of `*root*`, then a subject `*pro*` is added with the verb's features for number and person.
- If the verb has no subject and it is *not* a child of `*root*`, then a subject `*trace*` is added with the verb's features for number and person.

Finally, verbs are fully depassivized as follows: if a verb has the feature `(:voice passive)`, then `:subj` is changed to `:obj`, `:obj` to `:obj2` and `*pro*` is added as `:subj`.

4 Translation

The Spanish-English dictionary was constructed from multiple resources:

1. The lexicon of a Spanish Kimmo-based morphological analyzer that contained English glosses (Dorr, 1993).
2. Spanish-English word-lists from *freedict.com*, *spanish.about.com* and the web site of the freely available multilingual dictionary Ergane.[1].
3. A Spanish-English word list of abbreviations extracted from a part of the UN parallel corpus (none of the testing set used later was included).

```
abandonar V abandon/desert/forsake/leave/quit
abandonar V cede/give_in/give_up/give_way/relinquish/yield
cesin N abandonment
abandonado AJ forlorn/abandoned/abandonee
abandonamiento N indulgence
abandono N renunciation/dereliction/failing
abdicacin N abandonment/job/task
deber V owe/should::AUX/must::AUX/have_to::AUX
desamparado AJ abandonee/helpless
desamparar V forsake/abandon
descontar V depreciate/reduce/abandon_ship/cash_up/derate
descuidar V abandon/neglect

tense:impf FEAT tense:past
tense:pret FEAT tense:past
```

Figure 2: A Sample from the Matador Spanish-English Dictionary

The structure of the translation dictionary is a three-column file pairing a single Spanish lexeme with a POS and a forward-slash separated list of English lexemes. Components of multi-word words are separated with an underscore (see Figure 2). By default, a translation is assumed to affect the lexical choice but not the POS of the translated word. If this is not the case, the new POS is indicated by following the gloss with the marker "`::`" (e.g. *deber* in Figure 2). Features are specified as `<feature>:<value>` pairs with the special "POS" `FEAT`. In Figure 2, the Spanish imperfect tense feature is translated as the English past tense.

Overall, the dictionary contains 24,278 spanish lexemes with 50,606 word-POS-gloss triples (about 1.86 gloss per word-POS pair and 2.08 gloss per word). Almost half of the entries are nouns (48%). Adjectives make up 20% and verbs 18%. Proper nouns and Adverbs are 4% and 3% respectively. There are over 900 words (3%) with unknown part of speech.

Translation is accomplished through a simple word replacement algorithm. Matching is done using the word and POS. If the word-POS pair is not available, the translation algorithm attempts to back off to a union of all the translations of the word for all available parts of speech.

5 Exerge: Expansive Rich Generation for English

Exerge is a reusable GHMT generation component for translating from other languages into English.

5.1 Exerge Resources

Exerge utilizes three symbolic and two statistical English resources. The symbolic resources include the word-class lexicon, the categorial variation database (Habash and Dorr, 2003) and the syntactic thematic linking map. Statistical resources include a surface n-gram model and a structural n-gram model.

The first of the symbolic resources is the word-class lexicon, which defines verbs and prepositions in terms of their subcategorization frames and lexical conceptual primitives. A single verb or preposition can have multiple entries for each of its senses. For example, among other entries, run_1 as in ($John_{agent}$ $ran_{cause-go_{identificational}}$ $store_{theme}$) is distinguished from run_2 as in ($John_{theme}$ $ran_{go_{locational}}$). Second, the categorial-variation lexicon relates words to their categorial variants. For example, $hunger_V$, $hunger_N$ and $hungry_{AJ}$ are clustered together. So are $cross_V$ and $across_P$; and $stab_V$ and $stab_N$. The third symbolic resource is the syntactic-thematic linking map, which relates syntactic relations (such as subject and object) and prepositions to the thematic roles they can assign. For example, while a subject can take on just about any thematic role, an indirect object is typically a *goal*, *source* or *benefactor*. Prepositions can be more specific. For example, *toward* typically marks a *location* or a *goal*, but never a *source*.

In addition to a surface uni- and bi-gram model of English, a structural n-gram language model is used in Exerge. The structural n-gram model characterizes the relationship between words in a dependency representation of a sentence without taking into account the overall structure at the phrase level. This model is very useful for making lexical selection choices dependent on long distance relations not captured by surface n-gram model.

[1] http://download.travlang.com/Ergane/

5.2 Exerge Sub-modules

Exerge consists of seven steps (Figure 1). The first five are responsible for lexical and structural selection and the last two are responsible for linearization. Initially, the source language syntactic dependency, now with target lexemes, is normalized and restructured into a syntactico-thematic dependency format. The thematic roles are then determined in the thematic linking step. The syntax-thematic linking is achieved through the use of thematic grids associated with English (verbal) head nodes together with the syntactic-thematic linking map. This step is a *loose* linking step that does not enforce the subcategorization-frame ordering or preposition specification. This looseness is important for linking from unknown non-English subcategorization frames.

This is followed by structural expansion which explores conflated and inflated variations of the thematic dependency. Conflation is handled by examining all verb-argument pairs (V_{head},Arg) for *conflatability*. For example, in *John put salt on the butter*, *to put salt on* can be conflated as *to salt* but *to put on butter* cannot be conflated into *to butter*. The thematic relation between the argument and its head together with other lexical semantic features constrain this structural expansion. Head Swapping is restricted through a similar process that examines head-modifier pairs for *swappability*. The fourth step maps the thematic dependency to a target syntactic dependency. Syntactic positions are assigned to thematic roles using the verb class subcategorization frames and argument category specifications. The first four steps are all symbolically driven.

The fifth step is the first statistical component where dependency bigram statistics are used for lexical selection. The input to this step is an ambiguous syntactic dependency and the output is non-ambiguous. This step prunes ambiguous nodes using dependency bigram statistics. The purpose of this step is to constrain the overgeneration of the previous steps using a language model that is based on structural relations between lexemes. This is different from the last step (Statistical Ranking) in three ways: (1) it is structural not word-order-based, (2) it is based on lexemes not final surface forms, and (3) its effect is only seen on lexical selection whereas the n-gram statistical ranking determines both lexical selection and linearization.

Next is the linearization step, where a rule-based grammar is used to create a word lattice that encodes the different possible realizations of the sentence. The grammar is implemented using the linearization engine oxyGen (Habash, 2000). Finally, the word lattice is converted into a Halogen-compatible forest to be ranked with Halogen's Statistical Forest Ranker (Langkilde, 2000). Further details on generation in GHMT are provided in (Habash, 2002) and (Habash and Dorr, 2002).

6 Evaluation

This section presents an evaluation of the performance of Matador as a complete system.[2] The goal of this evaluation is to determine the output quality and robustness of Matador. For purposes of comparison, four systems are evaluated using test sets from three corpora with different genre.

The evaluation metric used is Bleu (BiLingual Evaluation Understudy) (Papineni et al., 2001). Bleu is a method of automatic translation evaluation that is quick, inexpensive and language independent. The Bleu score is basically an N-gram precision variation calculated as the ratio of the number of N-gram sequences in the generated string that appear in the reference (gold standard) string to the total number of N-gram sequences in the generated string. Bleu is used with 1 to 4-grams and without case sensitivity.[3]

Four systems are evaluated:

- Gisting (GIST): This simple approach is basically word-to-word translation (Resnik, 1997). A dictionary of 391,026 Spanish surface to English lexeme pairs is used with a unigram language model to resolve any ambiguity. This system is considered the baseline.
- Systran (SYST): This is a commercial Transfer-based (purely symbolic) MT system. The version used is Systran Spanish-English Professional edition with four translation glossaries (Political Science, Military Science, Legal and Business/Economics).[4] Systran's Spanish-English has been developed over several hundred person-years and is considered here the industry standard of Spanish-English MT.
- IBM Model 4 (IBM4): This is a primarily statistical MT system (Brown et al., 1993). The translation model was trained using Giza (Al-Onaizan et al., 1999) on 50,000 Spanish-English sentence pairs from the UN Spanish-English corpus (Graff, 1994). Simple tokenization was used and consisted of down-casing all words and separating all punctuation marks. The language model is built from the English side of the training data in addition to 450,000 sentences from the English side of the Arabic-English UN corpus (Jinxi, 2002). Decoding is done using ISI ReWrite Decoder (Germann and Marcu, 2000).[5]
- Matador (MTDR): All of the system's modules described earlier are used. The Structural n-gram language model was created using 127,000 parsed sentences from the English UN corpus covering over 3 million words. The model is limited to bigrams. The parsing was done using Connexor's English parser. SN-gram pruning is used only for lexical selection within dependencies. The parameters for conflation and inflation are set to allow a maximum of 10 variants per dependency. The surface N-gram language model used in ranking is the same as that used in the IBM4 system described above.

[2] The description and results of several intrinsic evaluations of specific Matador components will be presented in separate publication.

[3] Throughout this paper, Bleu scores are presented multiplied by 100.

[4] http://www.systransoft.com/

[5] Using fast greedy decoding, Model 4 translation and bigrams language modeling.

The Halogen ranking scheme used is bigrams with length normalization.

Three blind test sets are evaluated: (1) 2,000 sentences from the UN Spanish-English corpus (Graff, 1994); (2) 2,000 sentences from the FBIS Spanish-English corpus;[6] and (3) 1,000 sentences from the Bible. Each Spanish sentence had one English (gold standard) translation that was used as the Bleu reference. The one-reference behavior of Bleu is not optimal, but, unfortunately, there are no Spanish-to-multiple-English parallel corpora similar to the Chinese-English multiple translation corpus (LDC, 2002).

The three corpora, UN, FBIS and Bible, were selected to cover a wide range of genre to examine the behavior of the evaluated systems under different conditions. This is important since poverty of resources forces systems to be trained or built using whatever resources are available, which may not necessarily be the same as what needs to be translated. The results of the overall evaluation are shown in Table 1 and Figure 3.

The evaluation shows that although MTDR scores lower than IBM4 on the corpus where all language models were trained (UN), MTDR has a higher degree of robustness and scores higher when tested on text with new genre (Bible). SYST and GIST are the best and worst respectively for all corpora.[7]

As for runtime, GIST was the fastest, finishing all 2,000 sentences in the UN corpus in less than 16 seconds. This is followed by SYST (90 seconds[8]), IBM4 (8,495 seconds) and finally MTDR (14,155 seconds).[9]

The rest of this section compares the behavior of MTDR against IBM4 since SYST and GIST's performance make them excellent upper and lower bounds—score-wise and also by the amount of work needed to create them. The result comparison will focus on four aspects of the output: lexical choice, information loss, grammaticality and translation-divergence handling.

6.1 Lexical Choice

One problem with Bleu scoring when only a single reference is used is that lexical choice becomes more important for evaluation than other criteria that are dependent on the correct lexical choice. For example, the generation of a synonym of a word in the correct relative word order to another word scores less than the correct lexical choice for the two words in the wrong relative order. Moreover, all "incorrect" lexical choices are treated equally regardless of how close or different the chosen words are to the correct words. Even morphologically related words are not considered correct.

In the following example, the lexical choice is dangerously misleading yet the same basic word-mismatch penalty is applied:

(6) SP: Se **instalaron** tres nuevos mercados rurales.
EN: Three new rural markets were **established**.
IBM4: **minefields** three new rural markets .
MTDR: Three new rural markets were **installed**.

In example (6), the verb *instalar (establish)* was translated in IBM4 as *minefield*. The MTDR translation, *install* receive a penalty equal to that received by *minefield*.

This example, from the UN test data where IBM4 scored high, exemplifies an interesting pattern of behavior in statistical MT systems. IBM4 did either extremely well or extremely poorly at lexical choice. In some cases in the UN corpus, it generated almost perfect phrases, which it most likely saw in the training data. Its performance deteriorated significantly in the other two test sets. When IBM4 had to deal with a previously unseen word or sequence, it randomly picked words for translation and, in some cases, dangerously changed the meaning.

6.2 Information Loss

IBM4 consistently lost parts of the translated sentences. Other systems experienced occasional loss of information, too, but not at the same rate. IBM4's sentence length is on average 6% shorter than the gold standard. SYST and GIST both are 9% longer and MTDR is only 4% longer. The following is an example of typical loss of information in IBM4:

(7) SP: El daño **causado** al pueblo de Sud**áfrica** jamás debe subestimarse.
EN: The damage **caused** to the people of his country should never be underestimated.
IBM4: the damage the people of south must never underestimated .
MTDR: Never the **causado** damage to the people of South **Africa** should be underestimated.

In this example, the words *causado (caused)* and *África (Africa)* do not appear at all in IBM4. MTDR fails to translate *causado* correctly, but it is generated nonetheless. The missing preposition *to* in IBM4 (*the damage the people*) can cause an erroneous reading where *the people* are the agent of *underestimating*.

6.3 Grammaticality

MTDR handles the translation of linguistic features such as tense and pro-drop restoration much better than IBM4 does. The following example was observed in the UN test set:

(8) SP: Sin embargo, no **se suministró** información concreta respecto de esos casos.
EN: However, no specific information on the cases involved **was provided**.
IBM4: however , not **be provided** specific information on those cases .
MTDR: Specific information **was** not **provided** without embargo proportion of cases.

In example (8), two mistakes appear in IBM4: the auxiliary *be* is not conjugated correctly and the subject *specific information* appears after the verb rather than be-

[6] The U.S.Foreign Broadcast Information Service (FBIS).

[7] All Systems ran successfully on all sentences except that MTDR failed on a total of 21 sentences out of all 5,000 (0.42%). Failure happened exclusively at the last stage, in Halogen statistical ranking.

[8] This run was done on a Pentium 4 PC with 1.7Ghz and 512MB of memory

[9] Except for SYST, all other systems ran on a SparcIII, with 750Mhz and 1GB of memory.

Table 1: Overall Evaluation Results

	UN	FBIS	Bible
SYST	24.66 +/- 1.09	20.39 +/- 0.70	13.52 +/- 0.8
IBM4	24.78 +/- 1.21	11.76 +/- 0.55	4.71 +/- 0.66
MTDR	18.01 +/- 1.00	10.42 +/- 0.57	7.29 +/- 0.83
GIST	8.17 +/- 0.81	3.40 +/- 0.61	2.40 +/- 0.50

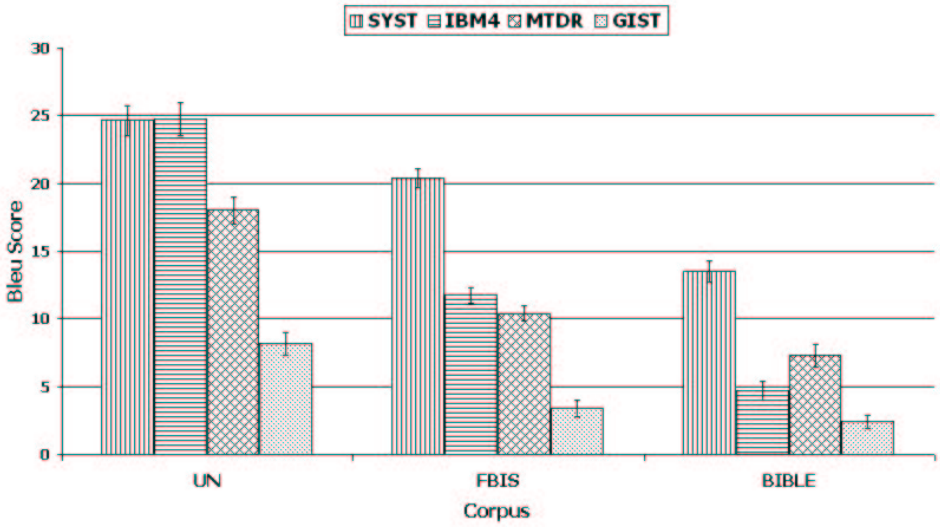

Figure 3: Overall Evaluation Results

fore it. In this case IBM4 was able to translate the complex Spanish passive verb into a passive English verb, but not moving the subject can cause a misunderstanding of whether *specific information* is really the subject or object of the verb *provide*.

Although IBM4 produced better lexical choice, the ungrammatical nature of the sentence structure can mislead readers about its content.

Bleu cannot capture syntactic long range phenomena that are spread over more than a 4-gram. To measure the "grammaticality" of the different systems' output, samples of 100 sentences from each the three tested corpora and their gold standard references were parsed using Connexor's English parser. The Spanish input was also parsed using Connexor's Spanish parser. The goal of this experiment is to determine the correctness of the tested systems output using parser decisions. Different outputs can be correct translations yet have radically different parses. Therefore, two specific phenomena that can be easily evaluated and that reflect complex long range parsing choices are focused on: verb determination and pro-drop restoration.

6.3.1 Verb Determination

Determining that a word is a verb depends on satisfying subcategorization specifications of that verb, such as presence of subject and object, etc. The number of words per sentence observed as verbs is calculated for all sample sentences (see Table 2).

Table 2 highlights the fact that the number of words observed as verbs in IBM4 sentences is radically smaller

Table 2: Verbs Per Sentence

	UN	FBIS	Bible
English	1.37	2.07	2.14
Spanish	1.49	2.10	2.23
GIST	1.72	2.25	2.08
IBM4	1.31	1.61	1.33
SYST	1.56	2.11	2.05
MTDR	1.46	2.08	2.20

than all other systems and the gold standard. It is an outlier for all three test sets being on average 77% smaller than the next closest value. This is even true for the UN and FBIS test sets where IBM4 scored a higher Bleu score than MTDR. This implies that sentence structure in IBM4 is consistently ungrammatical to a high degree.

6.3.2 Pro-drop Restoration

The restoration of dropped subjects when translating from languages like Spanish to English is very important to translation correctness. To determine how well the different systems accomplished this task, the ratio of "realized subjects" to all verbs is calculated. A subject is an argument with a `:subj` relation to a parent with a part of speech `V`. A realized subject is a subject that is not `*pro*` or `*trace*`. The results are shown in Table 3.

The first two rows in Table 3 highlight the stark difference in subject realization between English and Spanish. Less than 30% of English subjects are not realized. Examples include subjects appearing in subordinate clauses

Table 3: Percentage of Realized Subjects

	UN	FBIS	Bible
English	70.07%	75.84%	76.17%
Spanish	31.54%	36.19%	34.53%
GIST	41.86%	41.78%	41.83%
IBM4	55.73%	54.04%	51.13%
SYST	69.23%	72.99%	77.56%
MTDR	70.55%	68.75%	75.91%

Table 4: Handling of Translation Divergences in MTDR and IBM4

	Incorrect	Possible	Correct
IBM4	13 (32.5%)	14 (35%)	13 (32.5%)
MTDR	11 (27.5%)	16 (40%)	13 (32.5%)

as `*trace*`: *I want to (*trace*=I graduate)*. The percentage of realized subjects in the output of MTDR and SYST is consistent with that of the English gold standard. GIST's output, as would be expected, is closer to Spanish. The output of IBM4 is quite in the middle of the spectrum, which suggests some amount of pro-drop restoration but only 50% of the time. This last percentage is calculated as the ratio of the difference between IBM4 and Spanish over the difference between the English gold standard and Spanish.

6.4 Translation Divergences

A sample of size 100 sentences was extracted randomly from the UN Corpus to determine how different systems handled divergences. In this sample set, 40 cases of divergences were identified. Forty percent of these divergences needed pragmatic knowledge. For example, *ofercer posibilidades de comercializacion (offer possibilities for commercialization)* is translated as *be marketable*; and *tratar de establecer la paz (try to establish peace)* is translated as *to seek peace*. The rest are in principle resolveable in Matador, but Bleu gave MTDR credit for four cases and it gave IBM4 credit for 7 cases. In many cases, Matador was unable to produce the correct output or a variant of it due to failures in different components.

If problems with syntax or semantically-related lexical choice are ignored, it is possible to classify the handling of translation divergences into three broad categories: correct, possible and incorrect. Correctly handled cases are those in which a perfect match or a slightly different match is found. For example, the generation of *should* where the reference gold standard chooses *must* is essentially correct although no Bleu credit is given. Possible cases are those perfectly understood cases in which no divergence handling takes place or a different kind of divergence handling takes place. For example, one system generated *to be object of attention* instead of the reference's *to receive attention*. Finally, incorrect cases are totally wrong cases that are not understandable. For example, the output *this is not dable* contains an untranslated Spanish word (*dable* meaning *possible*). The reference translation is *this is impossible*.

Table 4 displays the number of divergence cases that fall in each of these three categories for both IBM4 and MTDR. The classification was done by one bilingual speaker of English and Spanish. The results show that in terms of divergence handling, MTDR and IBM4 have a comparable overall performance.

There are instances where Matador produced odd or incorrect output due to inappropriate over-expansions. In the following example, *destroy totally* is incorrectly conflated into $total_V$:

(9) SP: Además, **destruyó totalmente** sus cultivos de subsistencia ...
EN: It had **totally destroyed** Samoa's staple crops ...
IBM4: furthermore , **destroyed completely** their crops subsistence ...
MTDR: Furthermore, it **totalled** their cultivations of subsistence ...

This error resulted from a wrong entry in the Catvar database linking the adverb *totally* to the verb *total*.

7 Conclusions and Future Work

This paper has presented Matador, a Spanish-English Generation-Heavy Hybrid Machine Translation system. An extensive evaluation of Matador shows it to have a higher degree of robustness and superior output quality, in terms of grammaticality and accuracy, when compared to a primarily statistical approach that requires a parallel corpus.

Future work includes improving the quality of different components such as the categorial variation database. Additionally, a Chinese-English GHMT that reuses the Exerge component of Matador is currently under-development and an Arabic-English version is being planned. Finally, a human translation-quality evaluation is planned to address some of the issues associated with automatic evaluation techniques.

Acknowledgments

This work has been supported, in part, by Army Research Lab Cooperative Agreement DAAD190320020, NSF CISE Research Infrastructure Award EIA0130422, and Office of Naval Research MURI Contract FCPO.810548265. I would like to thank Bonnie Dorr for support and advice.

References

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I. D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical Machine Translation. Technical report, JHU. http://citeseer.nj.nec.com/al-onaizan99statistical.html.

Brown, P., Pietra, S. D., Pietra, V. D., and Mercer, R. (1993). The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Dorr, B. J. (1993). *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA.

Dorr, B. J., Jordan, P. W., and Benoit, J. W. (1999). A Survey of Current Research in Machine Translation. In Zelkowitz, M., editor, *Advances in Computers, Vol. 49*, pages 1–68. Academic Press, London.

Garcia, E. C. (1975). *The Role of Theory in Linguistic Analysis: The Spanish Pronoun System*. North-Holland Linguistic Series 19.

Germann, U. and Marcu, D. (2000). ISI ReWrite Decoder, Release 0.7.0b. University of Southern California. http://www.isi.edu/ germann/software/ReWrite-Decoder.

Graff, D. (1994). UN Parallel Text (Spanish-English), LDC Catalog No.: LDC94T4A. Linguistic Data Consortium, University of Pennsylvania.

Habash, N. (2000). oxyGen: A Language Independent Linearization Engine. In *Fourth Conference of the Association for Machine Translation in the Americas, AMTA-2000*, Cuernavaca, Mexico.

Habash, N. (2002). Generation-Heavy Machine Translation. In *Proceedings of the International Natural Language Generation Conference (INLG'02) Student Session*, New York.

Habash, N. and Dorr, B. J. (2002). Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. In *Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*, Tiburon, California.

Habash, N. and Dorr, B. J. (2003). A Categorial Variation Database for English. In *Proceedings of NAACL 2003*, Edmonton, Canada.

Jinxi, X. (2002). UN Parallel Text (Arabic-English), LDC Catalog No.: LDC2002E15. Linguistic Data Consortium, University of Pennsylvania.

Karlsson, F. (1990). Constraint Grammar as a Framework for Parsing Running Text. In *Proceedings of the 14th International Conference on Computational Linguistics (Coling'90)*.

Langkilde, I. (2000). Forest-based statistical sentence generation. In *Association for Computational Linguistics conference, North American chapter (NAACL'00)*.

LDC (2002). Multiple Translation Chinese-English Corpus, LDC Catalog No.: LDC2002T01. Linguistic Data Consortium, University of Pennsylvania.

Maldonado, R. (1988). Energetic reflexives in spanish. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society 14*, pages 153–165.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176(W0109-022), IBM Research Division, Yorktown Heights, NY.

Resnik, P. (1997). Evaluating multilingual gisting of web pages. Presentation at the AAAI Symposium on Natural Language Processing for the World Wide Web, Stanford, CA.

Tapanainen, P. and Jarvinen, T. (1997). A non-projective dependency parser. In *5th Conference on Applied Natural Language Processing / Association for Computational Linguistics*, Washington, D.C.