System Demonstration
Matador: Spanish-English GHMT

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

Abstract

This paper presents the online demo of Matador, a large-scale Spanish-English machine translation system implemented following the Generation-heavy Hybrid Machine Translation (GHMT) approach.

1 Introduction

Matador is a Spanish-English machine translation (MT) system implemented following the Generation-heavy Hybrid approach to Machine Translation (GHMT) (Habash, 2002; Habash and Dorr, 2002). The focus of GHMT is addressing resource poverty in MT by exploiting symbolic and statistical target language resources in source-poor/target-rich language pairs. Expected source language resources include a syntactic parser and a simple one-to-many translation dictionary. No transfer rules, complex interlingual representations or parallel corpora are used. Rich target language symbolic resources such as word lexical semantics, categorial variations and subcategorization frames are used to overgenerate multiple structural variations from a target-glossed syntactic dependency representation of source language sentences. This symbolic overgeneration is constrained by multiple statistical target language 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 source language parser and translation dictionary).[1]

The basic intuition of the GHMT approach parallels the experience of most language learners whose lack of symmetrical knowledge impairs their ability to translate into their newly learned language but does not hinder them as much when translating from the foreign language into their native tongue (where they are assisted by rich resources).

The next section is an overview of Matador's components. This is followed by a discussion of Matador's online demo.[2]

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 source-language-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. 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 ambiguous sets 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.

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. Parsing in Matador is implemented using the off-the-shelf Connexor Spanish parser (Tapanainen and Jarvinen, 1997).

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 the dependency tree are expansively created using lexical semantic

[1] The fact that a parser is an expensive resource to obtain must been seen in relative terms to the backgrounds from which GHMT originated: (a) Interlingual MT, a far more demanding approach that requires very expensive resources on top of syntactic parsers, and (b) Statistical MT, which requires very large parallel corpora (possibly annotated) that are hard to obtain for most languages too.

[2] A more detailed discussion of Matador including an extensive evaluation has been submitted to this conference's main session.

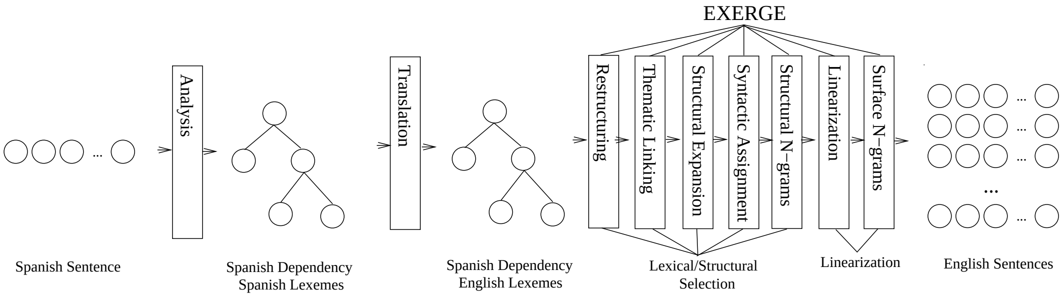


Figure 1: Matador: Spanish-English Generation-Heavy Machine Translation

information and other English-specific heavy resources. These expansions include *conflations* such as *to put butter* becoming *to butter* and *inflations* such as *to cross* becoming *to go across*. The following are only a few of these variants for (2):

```
(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: A rule-based grammar is used to create the word lattice. The grammar is implemented using the linearization engine oxyGen (Habash, 2000).

```
(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. Matador uses an off-the-shelf component for this step: Halogen's Statistical Forest Ranker (Langkilde, 2000). 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
```

3 Matador Online

Figure 2 presents Matador's online demo interface. The online demo is available at *http://clipdemos.umiacs.umd.edu/matador/*.

3.1 Text Input

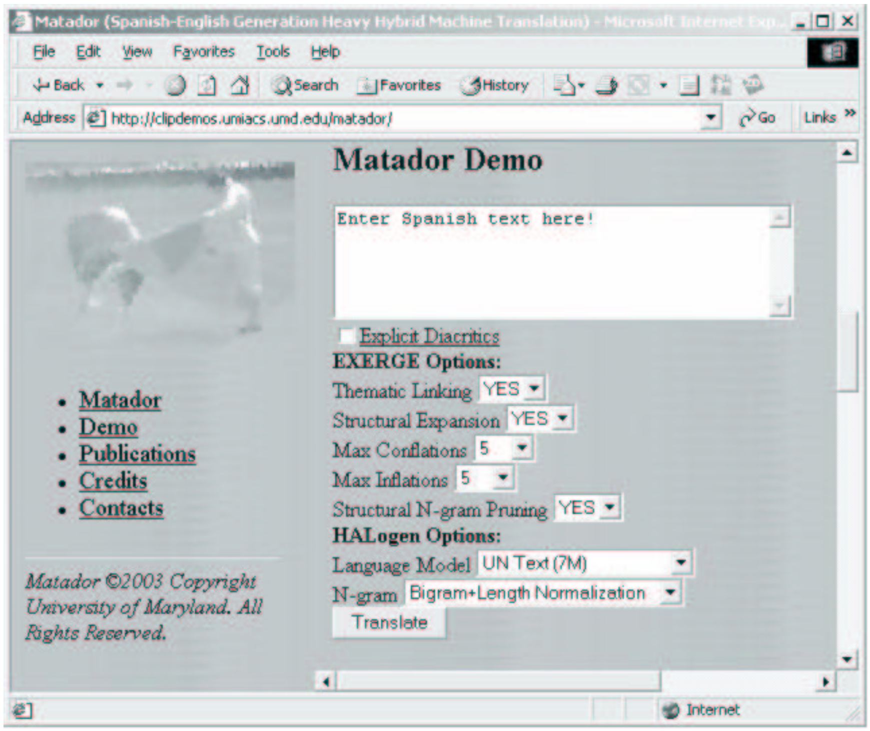
The interface allows cutting and pasting of Spanish text or typing text directly. If no Spanish text entry is provided, the user can specify *Explicit Diacritics* and use pairs of characters to specify diacritized characters in Spanish (see Table 1).

3.2 Exerge Options

The user can control the run mode of Exerge by specifying different parameters. The thematic linking, structural expansion and structural n-gram pruning are boolean parameters. The max conflation and max inflation parameters are integers. If thematic linking is turned off (option *No*), then Matador becomes a dependency gister. The Spanish parse tree is not modified and is treated as if it was an English parse tree in generation. All other parameters are ignored in such case.

If the structural expansion parameter is turned off and thematic linking is on, then the Spanish parse tree will be converted into a thematic dependency but no conflations or inflations are allowed. When the structural expansion parameter is turned on, the maximally allowable number of conflations or inflations is specified by the max conflation and max inflation parameters.

The structural n-gram pruning parameter controls whether structural n-grams are used to do lexical selection over the generated parse trees. The purpose of structural n-gram pruning 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 Halogen language model in that it is structural not word-order-based and in that it is based on lexemes not final surface forms. The structural n-gram language model was created using 127,000 parsed sentences from the English side of the Spanish-English (Graff, 1994) and Arabic-English (Jinxi, 2002) UN corpus covering over 3 million words. The model is limited to bigrams.



Matador (Spanish-English Generation Heavy Hybrid Machine Translation) - Microsoft Internet Exp...
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://clipdemos.umiacs.umd.edu/matador/
Go Links
Matador
Demo
Publications
Credits
Contacts
Translate
Internet

Matador Demo

Enter Spanish text here!

□ Explicit Diacritics

EXERGE Options:

Thematic Linking YES

Structural Expansion YES

Max Conflations 5

Max Inflations 5

Structural N-gram Pruning YES

HALogen Options:

Language Model UN Text (7M)

N-gram Bigram+Length Normalization

Matador ©2003 Copyright University of Maryland. All Rights Reserved.

Figure 2: Matador Online Demo

Table 1: Explicit Diacritics

Diacritized Character	**Explicitly Diacritized Character**	**Diacritized Character**	**Explicitly Diacritized Character**
á	a’	Á	A’
é	e’	É	E’
í	i’	Í	I’
ó	o’	Ó	O’
ú	u’	Ú	U’
ñ	n	Ñ	N
ü	u”	Ü	U”

3.3 Halogen Options

The interface allows the user to control two parameters to the surface N-gram ranking component implemented using Halogen (Langkilde-Geary, 2002). First, the language model can be selected as either the UN language model or Halogen News model. The UN language model is built from 500,000 sentences from the English side of the Spanish-English (Graff, 1994) and Arabic-English (Jinxi, 2002) UN corpus. The Halogen news model was provided as part of the Halogen package. It is trained on 250 million words from various news sources such as the Wall Street Journal and AP Newswire. Both language models were created using the CMU Statistical Toolkit (Clarkson and Rosenfeld, 1997). Besides the difference in genre and size, the UN language model was trained for unigrams and bigrams only, whereas the Halogen news language model has trigrams too. As a result, the size of Halogen News is 320M compared to only 7M for the UN language model.

The second parameter option is modeling scheme. There are three modes: unigram, bigram and trigram. These can be selected with or without length normalization.

3.4 Matador’s Online Output

The Matador interface shows all major intermediate steps in processing as part of its output. These include: parse, translation, expansions, lattice and the top ten sequences. Additionally, a time log is also printed out specifying the number of seconds spent in each component.

4 Conclusions and Future Work

The online demo of Matador and its various features and parameters have been presented. Future additions to the web interface include allowing URL specification for text input, providing access to more control parameters and creating more surface and structural language models to select from. The demo engine will also be used in building the Chinese-English and Arabic-English GHMT systems currently under development.

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

Clarkson, P. and Rosenfeld, R. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of EUROSPEECH 97*.

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.

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.

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

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

Langkilde-Geary, I. (2002). An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the International Natural Language Generation Conference (INLG-02)*, new York, New York.

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.