Comparing Corpus-based MT Approaches Using Restricted Resources

Monica Gavrila[1] Natalia Elita

University of Hamburg
[1]gavrila@informatik.uni-hamburg.de

International Workshop on Using Linguistic Information for Hybrid Machine Translation
18 November 2011

Outline

Introduction

The MT Systems

- The SMT System: Mb_SMT (A)
- The EBMT System: $Lin - EBMT^{REC+}$ (B)
- The Hybrid System: OpenMaTrEx (C)
- The On-line System: Google Translate (D)

The RoGER Corpus

The Experiments

- The Data
- The Results
- Analysis of the Results

Conclusions and Further Work

Framework

- (At least one) inflected language
- Lower-resourced language

Comparisons of MT System

- Comparing statistical MT (SMT), example-based MT (EBMT) and hybrid MT (EBMT-SMT) , when no additional linguistic information is added to the corpus.
  Can hybrid systems overtake the pure corpus-based MT (CBMT) approaches?
- Comparing SMT and EBMT, when part-of-speech (POS) information is added to the data.
  Does additional POS information bring improvement when small-sized data are involved? Which is the difference between SMT and EBMT?.

For a better overview we compare our results with the ones of an on-line MT system.

Language-pair: English-Romanian

The pure SMT system (**Mb_SMT**)

- follows the description of the baseline architecture given for the EMNLP 2011 6th Workshop on SMT[1];
- uses Moses[2], SRILM and GIZA++
- includes two changes: We use 3-grams and no tuning

[1] www.statmt.org/wmt11/baseline.html.

[2] www.statmt.org/moses/

The EBMT System: $Lin - EBMT^{REC+}$ (B)

$Lin - EBMT^{REC+}$:

- has been developed at the University of Hamburg;
- combines the linear EBMT approach with the template-based one;
- is based on surface-forms and uses no linguistic resources, with the exception of the parallel aligned corpus;
- contains all the three steps of an EBMT system: matching, alignment and recombination;

$Lin - EBMT^{REC+}$ Steps

The steps:

- training and test data are pre-processed.
- matching is based on surface-forms, focusing in finding recursively the longest common substrings.
- alignment information is extracted from the GIZA++ output of the **Mb_SMT** system.
- longest TL aligned subsequences are used further in the recombination step, which is based on 2-gram information and word-order constraints.
- ideas from the template-based EBMT approach are incorporated in the recombination step, by extracting and imposing several types of word-order constraints.

The Hybrid System: OpenMaTrEx (C)

- OpenMaTrEx is a free open-source (EBMT/hybrid MT) system based on the marker hypothesis.
- OpenMaTrEx can be run in two modes. We chose the one based on a Moses-based decoder (called MaTrEx[3]).
- Markers for English have already been contained in OpenMaTrEx.
- Markers for Romanian were created from scratch during the experiments presented in this paper, by using morpho-syntactic specifications from MULTEXT-East and Wikipedia.
- There are currently 366 Romanian and 307 English makers.

[3] `www.sf.net/projects/mosesdecoder/`.

The On-line System: Google Translate (D)

For comparison reasons we included an on-line MT System in our experiments: Google Translate (`translate.google.com`).

The RoGER Corpus

- developed at the University of Hamburg
- domain restricted (texts are from a users' manual of an electronic device);
- small-size (2333 sentences);
- parallel corpus, aligned at sentence level;
- Romanian (ro), English (en), German and Russian;
- manually compiled and verified;
- not annotated, diacritics are ignored, preprocessed text.

RoGER: Statistics

Feature	**English**	**Romanian**	**German**	**Russian**
No. tokens	26096	25850	27142	22383
Voc.* size	2012	3104	3031	3883
Voc. (*Frequency* > 2)	1231	1575	1698	1904

(**Voc.=vocabulary*).

Experimental Settings

English-Romanian: both directions of translation
2200 sentences for training, 133 for testing

1. Data with no annotation (I),
2. Data annotated with POS information (II): we annotated the corpus by means of the text processing web services described on `http://www.racai.ro/webservices/TextProcessing.aspx`.

Experimental Setting I

Data SL	No. of words	Voc. size	Average sentence length
en-ro			
Training	27889	2367	12.68
Test	1613	522	12.13
ro-en			
Training	28946	3349	13.16
Test	1649	659	12.40

Experimental Setting II

Data SL	No. of words	Voc. size	Average sentence length
en-ro			
Training	27816	2815	12.64
Test	1610	564	12.11
ro-en			
Training	28954	4133	13.16
Test	1651	735	12.41

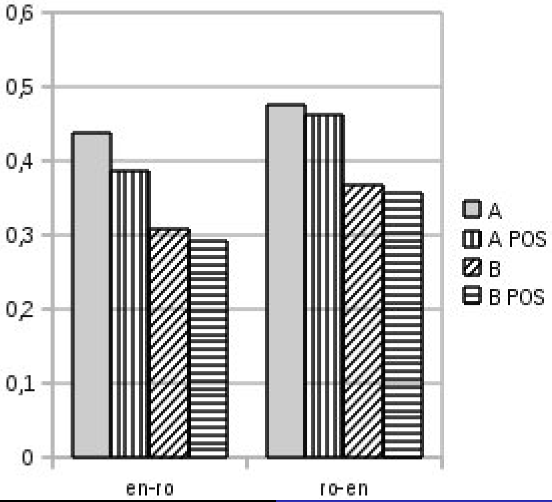
Experimental Setting I

Score	A	D	C	B
en-ro				
BLEU	0.4386	0.4782	0.3934	0.3085
NIST	6.5599	6.9334	5.9725	5.5322
ro-en				
BLEU	0.4765	0.5241	0.4428	0.3668
NIST	6.8022	7.4478	6.4124	6.2991

Experimental Setting II

Score	A	B
en-ro		
BLEU	0.3879	0.2916
NIST	5.8047	5.0893
ro-en		
BLEU	0.4618	0.3559
NIST	6.3533	6.0039

All the Results



Common Tokens

Desc.	Ref.	A	B
en-ro			
Total	495	490	466
CT	-	352 (71.11%)	302 (61.01%)
O. CT	-	343 (69.29%)	244 (49.29%)
en-ro and POS			
Total	490	472	480
CT	-	273 (55.71%)	257 (52.45%)
O. CT	-	267 (54.49%)	211 (43.06%)

I decided **to go** *home* **by** *bus.*

We **go to** *the theater* **by** *car.*

The sentences have 3 "*common tokens*" (CT) (*to*, *go*, *by*) and 2 "*ordered common tokens*" (OCT) (*go*, *by*).

Manual Evaluation

Evaluation	**A**	**B**
en-ro		
Adequacy	4.22	3.64
Fluency	4.08	3.44
en-ro and POS		
Adequacy	4.1	3.66
Fluency	3.74	3.3

Adequacy: 1=None, 2=Little, 3=Much, 4=Most, 5=All.

Fluency: 1=Incomprehensible, 2= Disfluent, 3=Non-native, 4=Good, 5=Flawless

Data Analysis

Out-of-vocabulary (OOV) Words and Sentences in the Training Data

Corpus	No. of OOV-Words (% from voc.* size)	Sentences in the corpus
en-ro		
Test	60 (11.49%)	37 (27.81%)
Test (POS)	74 (13.12%)	37 (27.81%)
ro-en		
Test	84 (12.75%)	34 (25.56%)
Test POS	116 (15.78%)	34 (25.56%)

Conclusions

- Several experiments for English and Romanian
- Different CBMT approaches and small-size data.
- Influence of POS information

- not always additional linguistic information improves the MT results
- combining different approaches does not always lead to better results
- training and test data themselves, the impact of additional information (such as increase of data sparseness) directly influence the translations

Further Work

Conclusion

For under-resourced language-pairs or lower-resourced domains it can be enough just the use of a pure SMT system.

Further work:

- further (manual) analysis is required
- run more tests with different language-pairs and corpora

Thank You!

Discussions

Questions? Suggestions? Remarks?