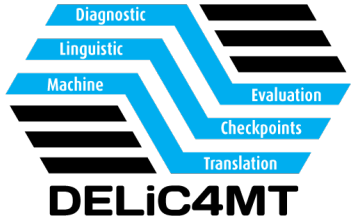


Diagnostic evaluation of MT with DELiC4MT

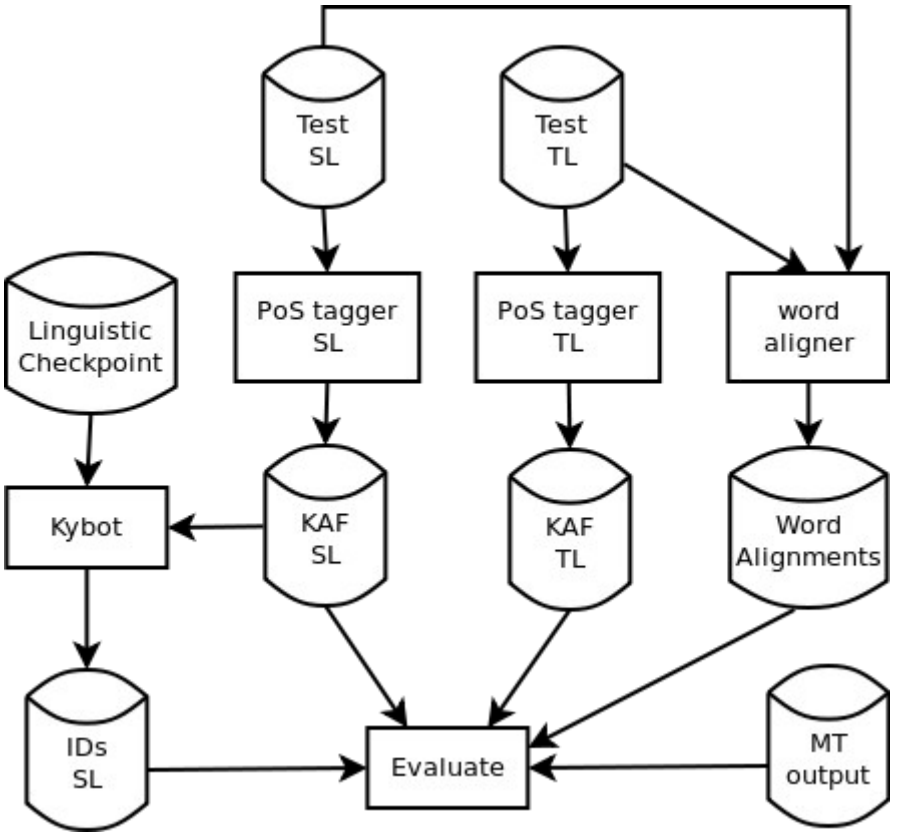
Final report

MT Marathon 2012
Edinburgh, 5th September 2012

Walid Aransa, Luong Ngoc Quang, Antonio Toral

Overview

- Automatic Diagonistic Evaluation on Linguistic Checkpoints



Ideas

- **Extend to a new language pair**
- Learn the linguistic phenomena
- Better filtering
- **Multiple references**
- **Optional parameter for alignment of MT output**
- Add precision-based metric
- Call any metric
- Metric considers not only forms but also lemmas and PoS
- Remove words from search once they are matched

Timeline/Progress

- Monday
  - Familiarisation with the tool trough tutorial use-case
    - Inner working, dependencies...
    - Improvement of tutorial and package
  - Brainstorming and selection of ideas
- Tuesday - Friday
  - Extending to a new language pair
  - Optional alignment for MT output
  - Multiple references

Extend to new language pair

- DELiC4MT currently supports EN, DE, FR, IT, NL, CY

- Add Arabic
  - Adapt PoS tagger
  - Identify and define linguistic phenomena for AR -> EN
  - Selection of test set and test evaluation of MT systems

Adding Support for Arabic

- List of added checkpoints:
  Adjective, Adjective+Noun, Adverb, preposition, preposition+Noun, Noun+Noun, Plural Noun, Number, pronoun, Verb.

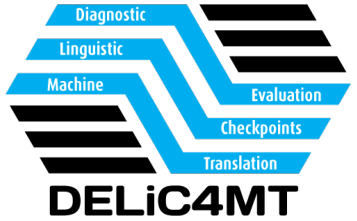
- Testing
  - Use Stanford PoS tagger (AR, EN)
  - Prepare KAF files
  - Prepare Kybot for each checkpoint
  - Evaluate

Sample Arabic Checkpoint

```
<?xml version="1.0" encoding="utf-8"?>

<Kybot id="kybot_n_a_ar">
  <variables>
    <var name="X" type="term" pos="NN*" />
    <var name="Y" type="term" pos="JJ*" />
  </variables>

<relations>
  <root span="X" />
  <rel span="Y" pivot="X" direction="following" immediate="true" />
</relations>

<events>
  <event eid="" target="$X/@tid" lemma="$X/@lemma" pos="$X/@pos"/>
  <role rid="" event="" target="$Y/@tid" lemma="$Y/@lemma" pos="$Y/@pos"/>
</events>

</Kybot>
```

Optional use of output source-target alignment

- Previous behaviour:
    1. N-grams matched against all the sentence in MT output
    2. This is still needed because MT src-trg alignment file not always available. So this is still applicable to any MT system (since we don't need alignment)
- New feature:
    1. Exploit MT alignment, if we have it!
    2. Keep the backward compatibility with old API interface and behaviour
    3. New arguments processing code to provide flexibility and easier maintainance of the code
    4. Added new switches to pass the alignment filename and other parameters
    5. Other enhancements: Added errors handling code



Multiple references

- Problem: 1 reference is not ideal, lexical variability, etc.
- Possible solutions:
  - Use lemmas, synonyms, paraphrases
  - Multiple references -> more expensive but more accurate

- Implementation of multiple references
  - Given time constraints and that all the info is in log files -> post-processing of log files for each reference
  - Read log for each ref
    - For each checkpoint instance, keep highest score
    - Get overall score

Evaluation scenarios

- 1 MT system
  - Which linguistic phenomena it translates best/worst?
  - Detect performance issues in terms of linguistic phenomena

- 2 (or more) MT systems
  - Comparative evaluation. Which system is better for which phenomenon? By how much?
  - Keep track of improvements wrt baseline

Results Comparative

	mt web1	mt web2	dif	
a0	0.5643	0.5253	-6.92%	
a1	0.5207	0.4793	-7.94%	
a2	0.5780	0.5382	-6.89%	-7.25%
n0	0.5606	0.5139	-8.32%	
n1	0.5230	0.4836	-7.53%	
n2	0.5612	0.5122	-8.73%	-8.19%
v0	0.5496	0.4972	-9.53%	
v1	0.5133	0.4722	-8.00%	
v2	0.5507	0.4976	-9.63%	**-9.05%**
r0	0.6627	0.6024	**-9.09%**	
r1	0.6500	0.6250	**-3.85%**	
r2	0.7229	0.6024	**-16.67%**	**-9.87%**
num0	0.5496	0.4972	-9.53%	
num1	0.5133	0.4722	-8.00%	
num2	0.5507	0.4976	-9.63%	**-9.05%**
pre0	0.7988	0.7720	-3.36%	
pre1	0.7766	0.7485	-3.62%	
pre2	0.8147	0.7752	-4.85%	**-3.94%**
pro0	0.8036	0.7619	-5.19%	
pro1	0.8035	0.7457	-7.19%	
pro2	0.7849	0.7500	-4.44%	-5.61%
in_N	0.5681	0.5135	-9.61%	
in_n	0.5344	0.4949	-7.39%	
in_n	0.5679	0.5245	-7.65%	-8.22%
n_a	0.4702	0.4331	-7.88%	
n_a	0.4403	0.4052	-7.97%	
n_a	0.4912	0.4481	-8.76%	-8.20%
			-7.71%	
bleu	0.2660	0.2130	-19.92%	
bleu	0.2240	0.1870	-16.52%	
bleu	0.2553	0.2025	-20.68%	
			-19.04%	

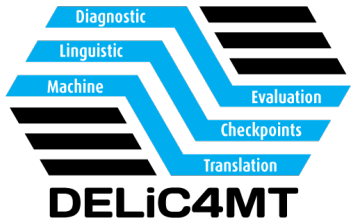
Diagnostic Linguistic Machine Evaluation Checkpoints Translation

DELiC4MT

Results
Multiple refs

	mt web1	mt web2	dif
a	0.7396	0.7075	-4.34%
n	0.7206	0.6751	-6.32%
v	0.7149	0.6616	**-7.45%**
r	0.7791	0.7442	-4.48%
num	0.7149	0.6616	**-7.45%**
pre	0.9091	0.8859	**-2.55%**
pro	0.9096	0.8763	**-3.65%**
in_n	0.7050	0.6493	**-7.90%**
n_a	0.6171	0.5677	**-8.01%**
			-5.79%

	mt web1	mt web2	dif	
a0	0.5643	0.5253	-6.92%	
a1	0.5207	0.4793	-7.94%	
a2	0.5780	0.5382	-6.89%	-7.25%
n0	0.5606	0.5139	-8.32%	
n1	0.5230	0.4836	-7.53%	
n2	0.5612	0.5122	-8.73%	-8.19%
v0	0.5496	0.4972	-9.53%	
v1	0.5133	0.4722	-8.00%	
v2	0.5507	0.4976	-9.63%	**-9.05%**
r0	0.6627	0.6024	**-9.09%**	
r1	0.6500	0.6250	**-3.85%**	
r2	0.7229	0.6024	**-16.67%**	**-9.87%**
num0	0.5496	0.4972	-9.53%	
num1	0.5133	0.4722	-8.00%	
num2	0.5507	0.4976	-9.63%	**-9.05%**
pre0	0.7988	0.7720	-3.36%	
pre1	0.7766	0.7485	-3.62%	
pre2	0.8147	0.7752	-4.85%	**-3.94%**
pro0	0.8036	0.7619	-5.19%	
pro1	0.8035	0.7457	-7.19%	
pro2	0.7849	0.7500	-4.44%	-5.61%
in_N	0.5681	0.5135	-9.61%	
in_n	0.5344	0.4949	-7.39%	
in_n	0.5679	0.5245	-7.65%	-8.22%
n_a	0.4702	0.4331	-7.88%	
n_a	0.4403	0.4052	-7.97%	
n_a	0.4912	0.4481	-8.76%	-8.20%
			-7.71%	
bleu	0.2660	0.2130	-19.92%	
bleu	0.2240	0.1870	-16.52%	
bleu	0.2553	0.2025	-20.68%	
			-19.04%	



Conclusions + Future

- Understanding about the tool
- Tasks conducted
  - New language pair
  - Optional alignment for MT
  - Multiple references

- Future work: wishlist is long :)

Thanks!
Questions?

Diagnostic evaluation of MT with DELiC4MT
Final report

MT Marathon 2012
Edinburgh, 5th September 2012

Walid Aransa, Luong Ngoc Quang, Antonio Toral