A Framework for Diagnostic Evaluation of MT Based on Linguistic Checkpoints

Sudip Kumar Naskar[†‡], **Antonio Toral**[†], **Federico Gaspari**[†]
NCLT[†] / CNGL[‡], School of Computing
Dublin City University
Dublin 9, Ireland
{snaskar,atoral,fgaspari}@computing.dcu.ie

Andy Way*
Applied Language Solutions
Huddersfield Road
Delph, UK
andy.way@appliedlanguage.com

Abstract

This paper describes an approach to the diagnostic evaluation of machine translation (MT) based on linguistic checkpoints, which can provide valuable information both to the developers and to the end-users of MT systems. We present a flexible framework and a new tool, DELiC4MT, for fine-grained diagnostic MT evaluation which can be extended to any language pair and applied to any evaluation target, once the phenomena of interest are covered by the linguistic analysis. As a case study, we evaluate the CoSyne[1] MT software against four leading web-based MT systems across a set of linguistic phenomena for three language pairs (from German, Italian and Dutch into English).

1 Introduction

The work presented in this paper was conducted in the CoSyne project, funded by the EU under the FP7 scheme. CoSyne involves seven partners: three academic institutions, University of Amsterdam (UvA, The Netherlands) as project coordinator, Fondazione Bruno Kessler (FBK, Italy) and Dublin City University (DCU, Ireland); one research organization, the Heidelberg Institute for Theoretical Studies (HITS, Germany); and, three end-user partners, Deutsche Welle (DW, Germany), Netherlands Institute of Sound and Vision (NISV, The Netherlands) and Wikimedia Foundation Netherlands (WMF). The CoSyne project aims at facilitating the synchronization of the contents of wiki sites across different languages using Machine Translation (MT), with the support of other components, e.g. textual entailment, document structure modeling and induction. The three language pairs covered in the first year of the project are German→English, Italian→English and Dutch→English (Toral et al., 2011). In the final year of the project, Turkish and Bulgarian will also be added, to show the adaptability of the system to less-resourced languages.

The aim of the work reported in this paper is to help the developers of the CoSyne MT system to improve the performance of the software, relying on the advice of the end-users on the basis of what they deem should be prioritized. Our focus is on providing means to identify classes of translation errors, devising an evaluation regime that is sufficiently fine-grained to capture the linguistic shortcomings of the MT system of particular concern to the end-users. This is meant not only to provide insights into the linguistic strengths and (especially) weaknesses of the MT component in the CoSyne system, but also to allow the MT developers to take corrective action by tweaking the parameters in the MT system, as appropriate. We believe that this work fills an important gap in the area of diagnostic evaluation.

[*] Work done while at CNGL, School of Computing, DCU.

[1] www.cosyne.eu/

The rest of the paper is organized as follows. In Section 2 we discuss previous research on diagnostic evaluation of MT. Section 3 describes the linguistic checkpoints-based diagnostic evaluation regime and the key components of the system architecture. Section 4 presents and analyzes the experimental results. Finally, Section 5 concludes and provides a roadmap for future work.

2 Related Work

Although the problem of providing diagnostic evaluation to improve MT systems has been occasionally addressed in the literature in the last five years or so, no widely accepted solutions seem to have emerged to date.

Vilar et al. (2006) present a framework to analyze manually the errors displayed by MT output which is based on a hierarchical structure covering five top-level classes: “Missing Words”, “Word Order”, “Incorrect Words”, “Unknown Words” and “Punctuation” errors. With this set-up they were able to identify the most important classes of errors for each language pair, e.g., incorrect verb tenses in translation from English into Spanish, and wrong word order in translation from Chinese into English. Farrús et al. (2011) carry out a manual error analysis on an MT system for Spanish—Catalan and classify the errors into linguistic levels (orthographic, morphological, lexical, semantic, and syntactic).

Popović et al. (2006) adopt a framework for the automatic analysis of MT errors based on the use of morpho-syntactic information, which shows that their linguistically-informed evaluation measures provide useful insights to understand the weaknesses of their MT system, while also indicating the best ways and methods to take remedial action.

Popović and Ney (2007) propose a method to zoom in on translation errors involving different Part-of-Speech (PoS) classes in the output. They apply this method to the estimation of inflectional errors and to the distribution of missing target-language words over PoS classes.

Following the hierarchy proposed in (Vilar et al., 2006), Popović and Burchardt (2011) present a tool that classifies errors into five categories. Parton and McKeown (2010) describe a novel algorithm to detect MT errors, focusing specifically on content words that are deleted. Xiong et al. (2010) attempt to automatically detect incorrect segments in MT output by training a classifier with a set of linguistic features.

Our aim for the work described in this paper was to come up with a sufficiently flexible and fine-grained regime for the diagnostic evaluation of MT that can be adapted to the needs of the end-users, while also providing meaningful information to the developers.

3 Linguistic Checkpoints-based Diagnostic Evaluation

In this section, we first give an overview of linguistic checkpoints and then detail the evaluation framework and the key components of the system.

3.1 Linguistic Checkpoints

A linguistic checkpoint can be defined as a linguistically-motivated unit, (e.g. an ambiguous word, a verb-object collocation, a POS-n-gram, a constituent, etc.) which is predefined in a linguistic taxonomy for diagnostic evaluation. Such a taxonomy is an inventory of linguistic phenomena of the source language that can present problems due to, for example, inherent ambiguity, or for translation into a specific target language, for instance because of syntactic divergence between the two languages involved in the translation process. The level of detail and the specific linguistic phenomena included in the taxonomy can vary, depending on what the developers and/or the end-users want to investigate as part of the diagnostic evaluation and on the number of aspects that they are interested in. Linguistic checkpoints form the basis of linguistic test suites which are the means by which the MT output is evaluated.

3.2 Diagnostic Evaluation Framework

This approach evaluates a system’s ability to handle various linguistic checkpoints. These were first proposed by Zhou et al. (2008), who developed Woodpecker,[2] a tool supporting diagnostic evaluation based on linguistic checkpoints. However, this tool has two important drawbacks. Firstly, language-dependent data for English–Chinese (the language pair considered in their paper) is hard-coded in the software, which means that adapting it

[2] http://research.microsoft.com/en-us/downloads/ad240799-a9a7-4a14-a556-d6a7c7919b4a/

to other language pairs is not straightforward. Secondly, its license (MSR-LA)[3] is quite restrictive, to the extent that researchers would not be able to publicly release their adaptations of the tool. For these reasons, we decided to implement a new tool supporting the functionality offered by Woodpecker, namely (i) automatic extraction of checkpoints using PoS taggers, word aligners and parsers, and (ii) n-gram-based evaluation of the matching checkpoints. The requirements we stipulated for this new tool include: (i) the code had to be well-organized and fully documented; (ii) creating new evaluation targets for any language pair has to be as easy as possible (no coding involved); and (iii) the tool should support different evaluation metrics.

Our novel tool, DELiC4MT (Diagnostic Evaluation using Linguistic Checkpoints For Machine Translation), [4] makes extensive use of already available components and representation standards. (i) It uses state-of-the-art PoS taggers and word aligners. Treetagger[5] and GIZA++[6] (Och and Ney, 2003), respectively, are used in the current version, although any similar tool could be used. (ii) It exploits the Travelling Object (TO) format, established in the FP7 PANACEA project,[7] to represent word alignment. This is an XML format for linguistic analysis (e.g., PoS tagging, parsing, etc.) and alignment (sentence/word) based on XCES.[8] Scripts to convert the output of well-established tools (GIZA++, Treetagger, etc.) are available. (iii) It uses the KYOTO Annotation Format (KAF) (Bosma et al., 2009), established in the FP7 KYOTO project, [9] to represent textual analysis. KAF represents each level of linguistic analysis based on ISO standards (i.e. MAF, SynAF, SemAF) and it is compatible with the Linguistic Annotation Framework (LAF) (Ide and Romary, 2003). (iv) It makes use of Kybots (Vossen et al., 2010), established in the FP7 KYOTO project, to define the evaluation targets (linguistic checkpoints). A Kybot profile can be thought of as a regular expression over elements and attributes in KAF documents.

The benefits of developing such a new tool include: (i) the end-user can easily create new evaluation targets or even adapt the tool to a new language pair, provided that the phenomena of interest are covered by the linguistic analysis available; (ii) the tool can work with any PoS tagger / word aligner, provided that their output can be converted to the KAF and TO formats, respectively; and (iii) it takes advantage of the outcomes of recently completed and ongoing FP7 projects.

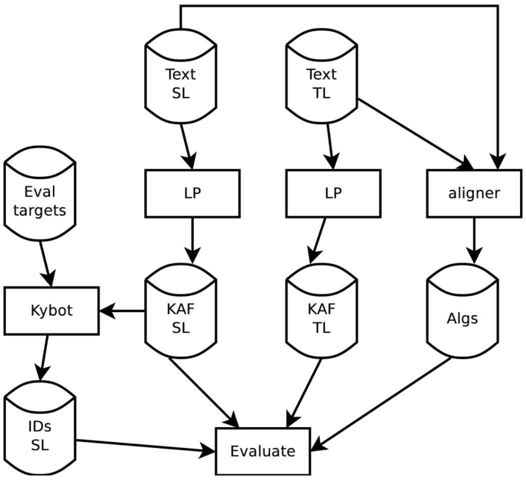

Figure 1. Linguistic checkpoints-based diagnostic evaluation scheme.

Figure 1 shows the architecture in which the evaluation based on linguistic checkpoints takes place. The gold standard of the test set in the source and target languages is processed by linguistic processors (PoS-tagging in the current version) and converted into KAF. Kybot profiles covering the different evaluation targets are run on the source KAF text, and the identifiers of the terms matched are stored. Ideally one would need gold standard manual alignments between the testset and the reference set, on which the diagnostic evaluation is crucially dependent. In the absence of such gold standard manual word alignment, however, automatic word aligners provide a good replacement. Finally, the evaluation module takes as input the identifiers from the Kybot, the KAF version of the test sets, the alignments and the output of an MT system, and outputs the result of the whole process. The following sections detail each of the stages involved.

[3] http://research.microsoft.com/en-us/downloads/ad240799-a9a7-4a14-a556-d6a7c7919b4a/MSR%20non-commercial%20license%20agreement.txt

[4] http://www.computing.dcu.ie/~atoral/delic4mt (under the license GPL-v3).

[5] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] http://code.google.com/p/giza-pp/

[7] http://panacea-lr.eu/

[8] http://www.xces.org/

[9] http://www.kyoto-project.eu/

3.3 Text Analysis and Conversion into KAF

In this first version of the diagnostic evaluation, the text is analyzed up to the morphological level using a PoS tagger. We have used Treetagger, as it is a state-of-the-art statistical PoS tagger and available for all the languages covered in the first year of the CoSyne project (Dutch, English, German and Italian).

Figure 2 shows sample KAF files produced from an English→Italian sentence pair (consisting of the Italian “È difficile rispondere” and its equivalent in English “That is hard to answer”):

```
<KAF>
 <text>
   [...]
   <wf wid="w962_1" sent="962" para="1">È</wf>
  <wf wid="w962_2" sent="962" para="1">difficile</wf>
  <wf wid="w962_3" sent="962" para="1">rispondere</wf>
   […]
 </text>
 <terms>
  [...]
 <term tid="t962_1" lemma="essere" pos="VER:pres">
   <span> <target id="w962_1"/> </span>
 </term>
 <term tid="t962_2" lemma="difficile" pos="ADJ">
   <span> <target id="w962_2"/> </span>
 </term>
 <term tid="t962_3" lemma="rispondere" pos="VER:infi">
   <span> <target id="w962_3"/> </span>
 </term>
  […]
 </terms>
 </KAF>
```

```
<KAF>
 <text>
   [...]
   <wf wid="w962_1" sent="962" para="1">That</wf>
   <wf wid="w962_2" sent="962" para="1">is</wf>
   <wf wid="w962_3" sent="962" para="1">hard</wf>
   <wf wid="w962_4" sent="962" para="1">to</wf>
   <wf wid="w962_5" sent="962" para="1">answer</wf>
  […]
  </text>
  <terms>
     <term tid="t962_1" lemma="that" pos="DT">
       <span> <target id="w962_1"/> </span>
     </term>
     <term tid="t962_2" lemma="be" pos="VBZ">
       <span> <target id="w962_2"/> </span>
     </term>
     <term tid="t962_3" lemma="hard" pos="JJ">
       <span> <target id="w962_3"/> </span>
     </term>
     <term tid="t962_4" lemma="to" pos="TO">
       <span> <target id="w962_4"/> </span>
     </term>
     <term tid="t962_5" lemma="answer" pos="VB">
       <span> <target id="w962_5"/> </span>
   </term>
   </terms>
</KAF>
```

Figure 2. Sample KAF files produced from an English→Italian sentence pair.

3.4 Word Alignment

To cater for the small testsets (1,000 sentence pairs for each language pair), which are far too small for a statistical word aligner to work, we append each testset to the Europarl corpus for that language pair, and then align them in order to obtain more reliable estimates. An advantage of using the TO format is that it allows us to compute union / intersection of alignments produced by different word alignment tools (e.g., GIZA++ and BerkeleyAligner[10]) to improve precision / recall of word alignment.

Figure 3 shows the GIZA++ word alignments for the sentence pair shown in Figure 1, converted into TO format;[11] source token [0] is aligned to target tokens [0,1], source token [1] to target token [2], and source token [2] to target tokens [3,4].

```
<cesAlign version="1.0">
  <cesHeader version="1.0">
    <profileDesc>
      <translations>
        <translation n="1" lang="it" trans.loc="test.it-en.it"
wsd="UTF-8"/>
        <translation n="2" lang="en" trans.loc="test.it-en.en"
wsd="UTF-8"/>
      </translations>
    </profileDesc>
  </cesHeader>
  <linkList>
  [...]
  <linkGrp domains=”s962 s962” targType=”t”>
    <link>
      <align xlink:href="#t0"/>
      <align xlink:href="#t0"/>
    </link>
    <link>
      <align xlink:href="#t0"/>
      <align xlink:href="#t1"/>
    </link>
    <link>
      <align xlink:href="#t1"/>
      <align xlink:href="#t2"/>
    </link>
    <link>
```

[10] http://code.google.com/p/berkeleyaligner/

[11] Note that identifiers in the alignment start from 0 while in the KAF files they do from 1.

```
      <align xlink:href="#t2"/>
      <align xlink:href="#t3"/>
    </link>
    <link>
      <align xlink:href="#t2"/>
      <align xlink:href="#t4"/>
    </link>
  </linkGrp>
  [...]
</cesAlign>
```

Figure 3. Word alignments in TO format for the English→Italian sentence pair shown in Figure 2.

3.5 Kybots

Kybots are used to extract the linguistic phenomena that are to be evaluated, which have been already established in the linguistic taxonomy for the source language.

```
<?xml version="1.0" encoding="utf-8"?>
<Kybot id="kybot_v_v">
  <variables>
    <var name="X" type="term" pos="VER*" />
    <var name="Y" type="term" pos="ADJ*" />
    <var name="Z" type="term" pos="VER*" />
  </variables>
<relations>
  <root span="X" />
  <rel span="Y" pivot="X" direction="following" immedi-
ate="true" />
  <rel span="Z" pivot="Y" direction="following" immedi-
ate="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" rtype="follows"/>
  <role rid="" event="" target="$Z/@tid"
lemma="$Z/@lemma" pos="$Z/@pos" rtype="follows"/>
</events>
</Kybot>
```

Figure 4. Kybot for the linguistic checkpoint "verb_adjective_verb".

The Kybot shown in figure 4, for example, is applied to Italian and extracts under the element "event" the term identifiers of those verbs that are immediately followed by an adjective, which in turn is immediately followed by another verb, like for example "è difficile rispondere". The equivalent tokens in the target corpus of those found by Kybots in the source language (in this example, "that is hard to answer") are obtained using the word alignments (presented in Section 3.4).

3.6 Diagnostic Evaluation

Diagnostic evaluation can be carried out at multiple levels: a checkpoint, a group of checkpoints or an entire linguistic taxonomy. For example, to measure the ability of an MT system to translate noun-noun compounds, all source sentences in the testset containing noun-noun checkpoints are selected using a Kybot that extracts these compounds. References for these noun-noun checkpoints are identified from the target side of the corresponding testset sentences through word alignment information. Then the system-generated translations for these sentences are matched against the references of the checkpoint under consideration.

To calculate the final score, we use a BLEU-style n-gram evaluation metric. We split each system-generated translation and reference for a checkpoint into a set of n-grams and compute the number of matching n-grams, and sum up the gains over all the n-grams as the score for this checkpoint, as in (Zhou et al., 2008). If the reference of the checkpoint is not consecutive, we use a wildcard character ("*"), which can be matched by any word sequence.

Given below are some examples for the Italian→English language pair to demonstrate the splitting and matching of n-grams.

- Consecutive checkpoint:

Checkpoint: "È difficile rispondere"
Reference: "that is hard to answer"
Candidate translation: "it is difficult to answer"
Matched n-grams: "is", "to", "answer", "to answer"

- Non-consecutive checkpoint:

Checkpoint: "È * rispondere"
Reference: "that is * answer"
Candidate translation: "it is difficult to answer"
Matched n-grams: "is", "answer", "is * answer"

When we calculate the recall of a set of checkpoints *C*, the references *r* of all checkpoints *c* in *C* (*c* can be a single checkpoint, a category, or a category group) are merged into one reference set *R*, on which the recall is obtained using equation (1).

$$R(C) = \frac{\sum_{r \in R} \sum_{n-gram \in r} match(n-gram)}{\sum_{r \in R} \sum_{n-gram \in r} count(n-gram)} \qquad (1)$$

Checkpoints	*Instances*	*Google*	*Bing*	*Freetranslation*	*Systran*	*CoSyne M12*
n	3709	**0.6037**^{b,c,d,e}	^{2}0.5462^{c,e}	^{5}0.5112	^{3}0.5384^{c,e}	^{4}0.5
v	1465	**0.4836**^{b,c,d,e}	^{2}0.4452^{c,d,e}	^{3}0.4089	^{4}0.4082	^{5}0.4054
a	1004	**0.5568**^{b,c,d,e}	^{2}0.5229^{c,d}	^{4}0.4841	^{5}0.4781	^{3}0.5010
r	553	**0.4232**^{c}	^{3}**0.3858**	^{5}0.3577	^{4}**0.3764**	^{2}**0.3989**
pre	1458	**0.6058**^{b,c,d}	^{3}0.5629	^{5}0.5355	^{4}0.5559	^{2}**0.5791**^{c}
pro	768	**0.5417**^{b,c,d,e}	^{2}0.4946^{c,d}	^{5}0.3616	^{4}0.4153^{c}	^{3}0.4570^{c}
appr_art_n	348	**0.4212**^{b,c,d}	^{3}0.3835^{c}	^{5}0.3453	^{4}0.3825^{c}	^{2}**0.4110**^{c}
art_adja_n	333	**0.4791**^{c,d}	^{3}**0.4461**^{c,d}	^{5}0.4022	^{4}0.4146	^{2}**0.4715**^{c,d}
BLEU		**0.2477**^{b,c,d,e}	^{2}0.2294^{c,d}	^{5}0.1657	^{4}0.1752^{c}	^{3}0.2052^{c,d}

Table 1. Results of the diagnostic evaluation for German→English.

Checkpoints	*Instances*	*Google*	*Bing*	*Freetranslation*	*Systran*	*CoSyne M12*
n	5955	**0.74**^{b,c,d,e}	^{3}0.6576^{c,d}	^{5}0.5859	^{4}0.5947	^{2}0.6817^{b,c,d}
v	2959	**0.6644**^{b,c,d,e}	^{3}0.5748^{c,d}	^{5}0.5119	^{4}0.5236	^{2}0.6037^{b,c,d}
a	2304	**0.7170**^{b,c,d,e}	^{3}0.6376^{c,d}	^{5}0.5602	^{4}0.5649	^{2}0.6651^{b,c,d}
r	837	**0.6658**^{b,c,d,e}	^{3}0.5696^{c,d}	^{5}0.5124	^{4}0.5150	^{2}0.5982^{c,d}
pre	3462	**0.7369**^{b,c,d,e}	^{3}0.6561^{c,d}	^{5}0.6149	^{4}0.6198	^{2}0.6921^{b,c,d}
pro	981	**0.7339**^{b,c,d,e}	^{3}0.6563^{c,d}	^{5}0.5776	^{4}0.5998	^{2}0.6885^{b,c,d}
polysemous	5725	**0.7062**^{b,c,d,e}	^{3}0.6231^{c,d}	^{5}0.5550	^{4}0.5638	^{2}0.6574^{b,c,d}
pos_seq3	1075	**0.5670**^{b,c,d,e}	^{3}0.4866^{c,d}	^{5}0.4071	^{4}0.4275^{c}	^{2}0.5185^{b,c,d}
pos_seq4	195	**0.5504**^{b,c,d,e}	^{3}0.4170	^{5}0.3929	^{4}0.4078	^{2}0.4986^{b,c,d}
n_di_n	773	**0.5743**^{b,c,d,e}	^{3}0.4991^{c,d}	^{5}0.4228	^{4}0.4415	^{2}0.5270^{c,d}
BLEU		**0.4235**^{b,c,d,e}	^{3}0.3106^{c,d}	^{5}0.1754	^{4}0.1840^{c}	^{2}0.3137^{c,d}

Table 2. Results of the diagnostic evaluation for Italian→English.

Checkpoints	*Instances*	*Google*	*Bing*	*Freetranslation*	*Systran*	*CoSyne M12*
n	7016	**0.6638**^{b,c,d,e}	^{2}0.6296^{c,d,e}	^{3}0.5615	^{4}0.5511	^{3}0.5615
v	2152	^{2}**0.4503**^{c,d,e}	**0.4508**^{c,d,e}	^{3}0.4261	^{5}0.4213	^{4}0.4246
a	1992	**0.7019**^{b,c,d,e}	^{2}0.6747^{c,d,e}	^{4}0.6315^{d}	^{5}0.6023	^{3}0.6481^{d}
r	534	^{2}**0.4725**	**0.4843**	^{4}**0.4510**	^{5}**0.4451**	^{3}**0.4627**
pre	3365	^{3}0.6760^{c,d}	**0.7042**^{a,c,d,e}	^{5}0.6148	^{4}0.6355^{c}	^{2}0.6766^{c,d}
pro	639	^{2}**0.4764**	**0.4878**^{e}	^{4}**0.4537**	^{3}**0.4602**	^{5}0.4439
BLEU		^{2}0.3330^{c,d,e}	**0.3347**^{c,d,e}	^{5}0.2456	^{4}0.2643^{c}	^{3}0.3223^{c,d}

Table 3. Results of the diagnostic evaluation for Dutch→English.

4 Results

We tested the linguistic checkpoints-based diagnostic evaluation tool on the following four free online MT systems: Google Translate,[12] Bing Translator,[13] Systran[14] and FreeTranslation[15] (the first two being statistical MT systems, and the others rule-based). We compared their performance against that of the CoSyne MT system in its month 12 implementation for the German→English, Italian→English and Dutch→English translation directions.

The test data for the three language pairs were taken from the news domain (Toral et al., 2011). As far as the linguistic checkpoints are concerned, we mainly considered PoS-based checkpoints (nouns (n), verbs (v), adjectives (a), adverbs (r), prepositions (pre) and pronouns (pro)), but any kind of combination of PoS-based checkpoints could be used in this diagnostic evaluation framework. Identification of such useful linguistic checkpoints requires expertise of the source language, and knowledge of which linguistic phenomena are potentially problematic when translating (e.g., due to syntactic divergence between the two specific languages).

[12] http://translate.google.com

[13] http://www.microsofttranslator.com

[14] http://www.systran.co.uk

[15] http://www.freetranslation.com

For Italian we consider some additional checkpoints, which include first of all polysemous words (9,000 polysemous lemmas are extracted from the Italian SIMPLE-CLIPS computational lexicon (Ruimy et al., 2002)). In addition, we look at frequent PoS sequences (the most frequent PoS 3-grams and PoS 4-grams in the Repubblica [16] and itWaC [17] corpora, i.e., "noun preposition-article noun" and "determiner (or preposition-article) noun preposition (or preposition-article) noun", respectively, and a problematic construction ('noun_*di*_noun'). [18] For German, we include as additional checkpoints the two most frequents PoS 3-grams found in the deWaC corpus;[17] namely, (i) preposition, article and noun (appr, art, n) and (ii) article, adjective and noun (art, adja, n).

The evaluation results are given in Tables 1-3. The best scores for each checkpoint, according to equation (1), and for BLEU are shown in bold in the tables (for ease of readability the rank of the other systems is shown in superscript before the actual scores). For example, a score of 0.5270 assigned to the CoSyne MT system for the checkpoint 'noun_*di*_noun' in Italian→English indicates that 52.7% of the n-grams (including skip n-grams) present in the reference for the 'noun_*di*_noun' checkpoint are found in the output produced by this system.

Results of statistical significance tests are also included to indicate the validity of the comparisons between the MT systems. Statistical significance is represented in the tables with characters written as superscripts after the scores. For each system and metric, a character *n* means that the current score is significantly better than the system in column *n* (P-value is set to 0.05 for checkpoints and 0.01 for BLEU). For example, *c,d* indicates that the current score is better than those obtained by the systems in the third and fourth columns.

As can be seen from Table 1, Google Translate obtains the best scores for all the checkpoints in German→English. Bing and CoSyne, the other two SMT systems considered in these evaluation experiments receive the 2nd and 3rd best scores for most of the checkpoints, with Bing often beating CoSyne. Interestingly, Systran and Freetranslation obtain the third best score for the checkpoints regarding nouns and verbs, respectively.

For Italian→English (Table 2) all the results are consistent, with Google, CoSyne and Bing obtaining the top 3 positions, in this order, for every single checkpoint considered.

However, for Dutch→English (Table 3), the results are more varied. Bing scores the highest for four checkpoints (verbs, adverbs, prepositions and pronouns) whereas Google does so for the remaining two (nouns and adjectives). This contrasts with the results obtained on automatic evaluation metrics where Bing is the clear winner (Toral et al., 2011). However, one should bear in mind that automatic evaluation metrics provide an overall score based on the entire test-set, whereas the linguistic checkpoint-based evaluation presented here is more fine-grained, which allows the evaluation of specific phenomena of interest to the MT developers and/or the end-users of the system.

5 Conclusions and Future Work

The diagnostic evaluation based on linguistic checkpoints presented in this paper affords the possibility to conduct fine-grained evaluation that is beyond the scope of state-of-the-art automatic MT evaluation metrics. The aim of the evaluation regime presented here is not to replace automatic MT evaluation metrics, but rather to supplement them. For example, for Dutch→English, the automatic evaluation metrics suggest that Freetranslation and Systran — the two rule-based MT systems — perform much worse than the three SMT systems. However, the diagnostic evaluation scores reveal that rule-based systems are not that far behind the statistical systems, at least with respect to the set of linguistic checkpoints included in this diagnostic evaluation. Considering the results of the diagnostic evaluation against those derived from the automatic metrics suggests that Google might not translate the content words so well for the Dutch→English language direction, but it must do a good job in word reordering, probably because it has a better language model for English than Bing.

[16] http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica

[17] http://wacky.sslmit.unibo.it/doku.php?id=corpora

[18] The Italian "di" roughly corresponding to the English preposition "of".

This knowledge derived from the diagnostic evaluation is crucial to the MT developers in determining which linguistic phenomena their MT systems are good at dealing with and, especially, where they fall behind. This is also useful to the end-users, who might decide to choose a particular MT system over another based on its capability to handle certain linguistic phenomena, e.g. envisaging the subsequent post-editing effort required.

As future developments for this diagnostic evaluation suite, we are working on the following tasks: (i) combining different word aligners to improve precision / recall of word alignment; (ii) supporting different evaluation metrics; (iii) developing complex evaluation metric(s); (iv) supporting evaluation targets with information up to the level of parsing; and (v) developing a complete suite of evaluation targets.

Acknowledgements

This work has been funded in part by the European Commission through the CoSyne project FP7-ICT-4-248531. We are also grateful to Aitor Soroa for his assistance on setting up and using Kybots.

References

Bosma, W., P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. KAF: a generic semantic annotation format. *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL2009), Workshop on Semantic Annotation*. Pisa, Italy, pp. 145-152.

Farrús, M., M.R. Costa-jussà, J.B. Mariño, M. Poch, A. Hernández, C. Henríquez and J.A.R. Fonollosa. 2011. Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan–Spanish language pair. *Language Resources and Evaluation* 45(2), pp. 181-208.

Ide, N., and L. Romary. 2003. Outline of the international standard Linguistic Annotation Framework. *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, Sapporo, Japan, pp. 1–5.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29 (1), pp.19-51.

Parton, K., and K. McKeown. 2010. MT error detection for cross-lingual question answering. *Proceedings of COLING 2010*, Beijing, China, pp. 946-954.

Popović, M., and A. Burchardt. 2011. From human to automatic error classification for machine translation output. *Proceedings of EAMT 2011*, Leuven, Belgium; pp. 265-272.

Popović, M., A. de Gispert, D. Gupta, P. Lambert, H. Ney, J.B. Mariño, M. Federico, and R. Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. *HLT-NAACL 2006: Proceedings of the Workshop on Statistical Machine Translation,* New York, NY, USA, pp. 1-6.

Popović, M., and H. Ney. 2007. Word error rates: decomposition over POS classes and applications for error analysis. *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic; pp. 48-55.

Ruimy, N., M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari and A. Zampolli. 2002. Clips, a multi-level Italian computational lexicon: A glimpse to data. *Proceedings of LREC 2002*, Las Palmas de Gran Canaria, Spain, pp. 792-799.

Toral, A., F. Gaspari, S.K. Naskar, and A. Way. 2011. A Comparative Evaluation of Research vs. Online MT Systems. *Proceedings of EAMT 2011*, Leuven, Belgium, pp. 13-20.

Vilar, D., J. Xu, L.F. D'Haro, and H. Ney. 2006. Error analysis of statistical machine translation output. *Proceedings of LREC-2006*. Genoa, Italy, pp. 697-702.

Vossen, P., G. Rigau, E. Agirre, A. Soroa, M. Monachini, and R. Bartolini. 2010. KYOTO: an open platform for mining facts. *COLING-2010: Proceedings of the 6th international workshop on Ontologies and Lexical Resources (Ontolex 2010)*, Beijing, China, pp. 1-10.

Xiong, D., M. Zhang, and H. Li. 2010. Error detection for statistical machine translation using linguistic features. *Proceedings of ACL 2010*, Uppsala, Sweden, pp. 604-611.

Zhou, M., B. Wang, S. Liu, M. Li, D. Zhang, and T. Zhao. 2008. Diagnostic evaluation of machine translation systems using automatically constructed check-points. *Proceedings of COLING 2008*, Manchester, UK, pp. 1121-1128.