

# Diagnostic Evaluation of Machine Translation Systems Using Automatically Constructed Linguistic Check-Points

Ming Zhou<sup>1</sup>, Bo Wang<sup>2</sup>, Shujie Liu<sup>2</sup>, Mu Li<sup>1</sup>, Dongdong Zhang<sup>1</sup>, Tiejun Zhao<sup>2</sup>

<sup>1</sup>Microsoft Research Asia  
Beijing, China

{mingzhou, muli, dozhang}  
@microsoft.com

<sup>2</sup>Harbin Institute of Technology  
Harbin, China

{bowang, Shujieliu, tjzhao}  
@mtlab.hit.edu.cn

## Abstract

We present a diagnostic evaluation platform which provides multi-factored evaluation based on automatically constructed check-points. A check-point is a linguistically motivated unit (e.g. an ambiguous word, a noun phrase, a verb~obj collocation, a prepositional phrase etc.), which are pre-defined in a linguistic taxonomy. We present a method that automatically extracts check-points from parallel sentences. By means of check-points, our method can monitor a MT system in translating important linguistic phenomena to provide diagnostic evaluation. The effectiveness of our approach for diagnostic evaluation is verified through experiments on various types of MT systems.

## 1 Introduction

Automatic MT evaluation is a crucial issue for MT system developers. The state-of-the-art methods for automatic MT evaluation are using an n-gram based metric represented by BLEU (Papineni et al., 2002) and its variants. Ever since its invention, the BLEU score has been a widely accepted benchmark for MT system evaluation. Nevertheless, the research community has been aware of the deficiencies of the BLEU metric (Callison-Burch et al., 2006). For instance, BLEU fails to sufficiently capture the vitality of natural languages: all grams of a sentence are

treated equally ignoring their linguistic significance; only consecutive grams are considered ignoring the skipped grams of certain linguistic relations; candidate translation gets acknowledged only if it uses exactly the same lexicon as the reference ignoring the variation in lexical choice. Furthermore, BLEU is useful for optimizing and improving statistical MT systems but it has shown to be ineffective in comparing systems with different architectures (e.g., rule-based vs. phrase-based) (Callison-Burch et al., 2006).

Another common deficiency of the state-of-the-art evaluation approaches is that they cannot clearly inform MT developers on the detailed strengths and flaws of an MT system, and therefore there is no way for us to understand the capability of certain modules of an MT system, and the capability of translating certain kinds of language phenomena. For the purpose of system development, MT developers need a diagnostic evaluation approach to provide the feedback on the translation ability of an MT system with regard to various important linguistic phenomena.

We propose a novel diagnostic evaluation approach. Instead of assigning a general score to an MT system we evaluate the capability of the system in handling various important linguistic test cases called **Check-Points**. A check-point is a linguistically motivated unit, (e.g. an ambiguous word, a noun phrase, a verb~obj collocation, a prepositional phrase etc.) which are pre-defined in a linguistic taxonomy for diagnostic evaluation. The reference of a check-point is its corresponding part in the target sentence. The evaluation is performed by matching the candidate translation corresponding to the references of the check-points. The extraction of the check-points is an automatic process using word aligner and parsers. We control the noise of the word aligner and parsers within tolerable scope by selecting

---

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (<http://creativecommons.org/licenses/by-nc-sa/3.0/>). Some rights reserved.

reliable subset of the check-points and weighting the references with confidence.

The check-points of various kinds extracted in this way have shown to be effective in performing diagnostic evaluation of MT systems. In addition, scores of check-points are also approved to be useful to improve the ranking of MT systems as additional features at sentence level and document level.

The rest of the paper is structured in the following way: Section 2 gives the overview of the process of the diagnostic evaluation. Section 3 introduces the design of check-point taxonomy. Section 4 explains the details of construction of check-point database and the methods of reducing the noise of aligner and parsers. Section 5 explains the matching approach. In Section 6, we introduce the experiments on different MT systems to demonstrate the capability of the diagnostic evaluation. In Section 7, we show that the check-points can be used to improve the current ranking methods of MT systems. Section 8 compares our approach with related evaluation approaches. We conclude this work in Section 9.

## 2 Overview of Diagnostic Evaluation

In our implementation, we first build a check-point database encoded in XML by associating a test sentence with qualified check-points it contains. This process can be described as following steps:

- Collect a large amount of parallel sentences from the web or book collections.
- Parse the sentences of source language and target language.
- Perform the word alignments between each sentence pair.
- For each category of check-points, extract the check-points from the parsed sentence pairs.
- Determine the references of each check-point in source language based on the word alignment.

With the extracted check-point database, the diagnostic evaluation of an MT system is performed with the following steps:

- The test sentences are selected from the database based on the selected categories of check-points to be evaluated.
- For each check-point, we calculate the number of matched n-grams of the refer-

ences against the translated sentence of the MT system. The credit of the MT system in translating this check-point is obtained after necessary normalization.

- The credit of a category can be obtained by summing up the credits of all check-points of this category. Then the credit of an MT system can be obtained by summing up the credits of all categories.
- Finally, scores of system, category groups (e.g. Words), single category (e.g. Noun), and detail information of n-gram matching of each check-point are all provided to the developers to diagnose the MT system.

## 3 Linguistic Check-Point Taxonomy

The taxonomy of automatic diagnostic evaluation should be widely accepted so that the diagnostic results can be explained and shared with each other. We will also need to remove the sophisticated categories that are out of the capability of current NLP tools to recognize.

In light of this consideration, for Chinese-English machine translation, we adopted the manual taxonomy introduced by (Lv, 2000; Liu, 2002) and removed items that are beyond the capability of our parsers. The taxonomy includes typical check-points at word, phrase and sentence levels. Some examples of the representative check-points at different levels are provided below:

- Word level check-points:
  - a. Preposition word e.g., 于(in), 在(at)
  - b. Ambiguous word e.g., 打(play)
  - c. New word<sup>1</sup> e.g., 朋克(Punk)
- Phrase level check-points:
  - a. Collocation. e.g., 油炸-食品(fired – food)
  - b. Repetitive word combination. e.g., 看看(have a look)
  - c. Subjective-predicate phrase e.g., 他\*说, (he\*said)
- Sentence level check-points:
  - a. “BA” sentence<sup>2</sup>: 他把(BA)书拿走了。(He took away the book.)
  - b. “BEI” sentence<sup>3</sup>: 花瓶被(BEI)打碎了。(The vase was broken.)

---

<sup>1</sup> New words are the terms extracted from web which can be a name entity or popular words emerging recently.

<sup>2</sup> In a “BA” sentence, the object which normally follows the verb occurs preverbally, marked by word “BA”.

<sup>3</sup> “BEI” sentence is a kind of passive voice in Chinese marked by word “BEI”.

Our implementation of Chinese-English check-point taxonomy contains 22 categories and English-Chinese check-point taxonomy contains 20 categories. Table 1 and 2 show the two taxonomies. In practice, any tag in parsers (e.g. NP) can be easily added as new category.

Word level		
Ambiguous word	New word	Idiom
Noun	Verb	Adjective
Pronoun	Adverb	Preposition
Quantifier	Repetitive word	Collocation
Phrase level		
Subject-predicate phrase	Predicate-object phrase	Preposition-object phrase
Measure phrase	Location phrase	
Sentence level		
BA sentence	BEI sentence	SHI sentence
YOU sentence	Compound sentence	

Table 1: Chinese check-point taxonomy

Word level		
Noun	Verb (with Tense)	Modal verb
Adjective	Adverb	Pronoun
Preposition	Ambiguous word	Plurality
Possessive	Comparative & Superlative degree	
Phrase level		
Noun phrase	Verb phrase	Adjective phrase
Adverb phrase	Preposition phrase	
Sentence level		
Attribute clause	Adverbial clause	Noun clause
Hyperbaton		

Table 2: English check-point taxonomy

## 4 Construction of Check-Point Database

Given a bilingual corpus with word alignment, the construction of check-point database consists of following two steps. First, the information of pos-tag, dependency structure and constituent structure can be identified with parsers. Then check-points of different categories are identified. Check-points of word-level categories such as Chinese idiom and English ambiguous words are extracted with human-made dictionaries, and the check-points of New-Word are extracted with a new word list mined from the web. A set of human-made rules are employed to extract certain categories involving sentence types such as compound sentence.

Second, for a check-point, with the word alignment information, the corresponding target language portion is identified as the reference of this check-point. The following example illustrates the process of extracting check-points from a parallel sentence pair.

- A Chinese-English sentence pair:  
他们反对建立储备金。  
They opposed the building of reserve funds.
- Word segmentation and pos-tagging:  
他们/R 反对/V 建立/V 储备金/N . /P
- Parsing result (e.g. a dependency result):  
(SUB, 1/反对, 0/他们) (OBJ, 1/反对, 2/建立) (OBJ, 2/建立, 3/储备金)
- Word alignment:  
(1; 1); (2; 2); (3; 4); (4; 6,7);
- The check-points in table 3 are extracted:

Category	Check-point	Reference
New word	储备金	<i>reserve funds</i>
Ambiguous word	建立	<i>building</i>
Predicate – object phrase	建立储备金	<i>the building of reserve funds</i>
Subject-predicate phrase	他们反对	<i>They opposed</i>

Table 3: Example of check-point extraction

To extract the categories of check-points of different schema of syntactic analysis such as constitute structure and dependency structure, three parsers including a Chinese skeleton parser (a kind of dependency parser) (Zhou, 2000), Stanford statistical parser and Berkeley statistical parser (Klein 2003) are used to parse the Chinese and English sentences. As explained in next section, these multiple parsers are also used to select high confident check-points. To get word alignment, an existing tool GIZA++ (Och 2003) is used.

### 4.1 Reducing the Noise of the Parser

The reliability of the check-points mainly depends on the accuracy of the parsers. We can achieve high quality word level check-points with the state-of-the-art POS tagger (94% precision) and dictionaries of various purposes. For sentence level categories, the parser tags and manually compiled rules can also achieve 95% accuracy. For some kinds of categories at phrase level which parsers cannot produce high accuracy, we only select the check-points which can be identified by multiple parsers, that is, adopt the intersection of the parsers results. Table 4 shows the improvement brought by this approach. Column 1 and 2 shows the precision of 6 major types of phrases in Stanford and Berkeley parser. Column 3 shows the precision of intersection and column 4 shows the reduction of the number of check-points when adopting the intersection results. The test corpus is a part of

Penn Chinese Treebank which is not contained in the training corpus of two statistical parsers. (Klein 2003).

	Stf%	Brk%	Inter%	Tpts redu%
NP	87.37	86.03	95.83	17.06
VP	87.34	82.87	95.23	19.68
PP	90.60	88.56	96.00	11.50
QP	98.12	92.90	99.21	6.31
ADJP	91.95	90.87	96.41	10.20
ADVP	95.21	94.25	92.64	3.92

Table 4: Precision of parsers and their intersection (Stf is Stanford, Brk is Berkeley)

## 4.2 Alleviating the Impact of Alignment Noise

Except for sentence level check-points whose references are the whole sentences and New Word, Idiom check-points whose references are extracted from dictionary, the quality of the references are impacted by the alignment accuracy. To alleviate the noise of aligner we use the lexical dictionary to check the reliability of references. Suppose  $c$  is a check-point, for each reference  $c.r$  of  $c$  we calculate the dictionary matching degree  $DM(c.r)$  with the source side  $c.s$  of  $c$ :

$$DM(c.r) = \text{Max}(0.1, \frac{\text{CoCnt}(c.r, \text{Dic}(c.s))}{\text{WordCnt}(c.r)}) \quad (1)$$

Where  $\text{Dic}(x)$  is a word bag contains all words in the dictionary translations of each source word in  $x$ .  $\text{CoCnt}(x, y)$  is the count of the common words in  $x$  and  $y$ .  $\text{WordCnt}(x)$  is the count of words in  $x$ . Specially, if  $c.r$  is not obtained based on alignment  $DM(c.r)$  will be 1. Because the limitation of dictionary, a zero  $DM$  score not always means the reference is completely wrong, so we force the  $DM$  score to be not less than a minimum value (e.g. 0.1).  $DM$  score will further be used in evaluation in section 5.

To better understand the reliability of the references and explore whether increasing the number of check-points could also alleviate the impact of noise, we built two check-point databases from a human-aligned corpus (with 60,000 sentence pairs) and an automatically aligned corpus (using GIZA++) respectively and tested 10 different SMT systems<sup>4</sup> with them. The Spearman correlation is calculated between two ranked lists of the 10 evaluation results against the two data-

<sup>4</sup> These systems are derived from an in-house phrase based SMT engine with different parameter sets.

bases. A higher correlation score means that the impact of the mistakes in word alignment is weaker. The experiment is repeated on 6 subsets of the database with the size from 500 sentences to 16K sentences to check the impact of the corpus size.

At system level, high correlations are found at different corpus sizes. At category level, correlations are found to be low for some categories at small size and become higher at larger corpus size. The results indicate that the impact of the alignment quality can be ignored if the corpus size is at large scale. As the check-points can be extracted fully automatically, increasing the size of check-point database will not bring extra cost and efforts. Empirically, the proper scale is set to be 2000 or more sentences according to the Table 6.

	500	1K	2K	4K	8K	16K
<b>Ambiguous word</b>	0.98	0.98	0.98	0.98	0.96	0.98
<b>Noun</b>	0.93	0.99	0.99	0.89	0.8	0.86
<b>Verb</b>	0.97	0.97	0.99	0.99	0.95	0.92
<b>Adjective</b>	0.16	0.19	0.57	0.75	0.77	0.97
<b>Pronoun</b>	0.96	1	0.93	0.99	0.97	0.99
<b>Adverb</b>	0.38	0.77	0.8	0.96	0.72	0.84
<b>Preposition</b>	0.56	0.86	0.9	0.9	0.97	0.96
<b>Quantifier</b>	1	0.46	0.46	0.98	0.85	0.96
<b>Repetitive Word</b>	0.99	0.99	0.97	0.89	0.73	0.95
<b>Collocation</b>	0.42	0.77	0.77	0.77	0.73	0.88
<b>Subject-predicate phrase</b>	0.06	0.8	0.95	1	0.96	0.84
<b>Predicate-object phrase</b>	0.84	0.96	0.78	0.7	0.78	0.88
<b>Preposition-object phrase</b>	0.51	0.5	0.93	0.95	0.87	0.99
<b>Measure phrase</b>	0.91	0.67	0.95	0.95	0.87	0.97
<b>Location phrase</b>	0.62	0.54	0.55	0.55	0.85	0.89
<b>SYSTEM</b>	0.95	0.95	0.98	0.99	0.97	0.98

Table 6: Impact of word alignment at different sizes of test corpus.

## 5 Matching Check-Points for Evaluation

Evaluation can be carried out at multiple options: for certain linguistic category, a group of categories or entire taxonomy. For instance, in Chinese-English translation task, if a MT developer would like to know the ability to translate idiom, then a number of parallel sentences containing idiom check-points are selected from the database. Then the system translation sentences are matched to the references of the check-points of idioms.

To calculate the credit at different occasions of matching, similar to BLEU, we split each reference of a check-point into a set of n-grams and sum up the gains over all grams as the credit of this check-point. Especially, if the check-point is not consecutive we use a special token (e.g. “\*”) to represent a component which can be wildcard matched by any word sequence. We use the following examples to demonstrate the splitting and matching of grams.

- Consecutive check-point:  
 Check-point: 在打鼓  
 Reference: playing a drum  
 Candidate translation: He is playing a drum.  
 Matched n-grams: playing; a; drum; playing a; a drum; playing a drum
- Not consecutive check-point:  
 Check-point: 他们\*打  
 Reference: They\*playing  
 Candidate translation: They are playing cop per drum.  
 Matched n-grams: They; playing; They \* playing

Additionally, to match word inflections, 3 different options of matching granularity are defined as follows.

- Normal: matching with exact form.
- Lower-case: matching with lowercase.
- Stem: matching with the stem of the word.

For a check-point  $c$  and references set  $R$  of  $c$ , we select the  $r^*$  in  $R$  which matches the translation best based on formula (2).

$$r^* = \arg \max_{r \in R} (DM(r) \cdot \frac{\sum_{n\text{-gram} \in r} Match(n\text{-gram})}{\sum_{n\text{-gram}' \in r} Count(n\text{-gram}')} ) \quad (2)$$

When we calculate the recall of a set of check-points  $C$  ( $C$  can be a single check-point, a category or a category group),  $r^*$  of each check-point  $c$  in  $C$  are merged into one reference set  $R^*$  and the recall of  $C$  is obtained using formula (3) on  $R^*$ .

$$Re(C) = \frac{\sum_{r \in R^*} (DM(r) \cdot \sum_{n\text{-gram} \in r} Match(n\text{-gram}))}{\sum_{r' \in R^*} (DM(r') \cdot \sum_{n\text{-gram}' \in r'} Count(n\text{-gram}'))} \quad (3)$$

A penalty is also introduced to punish the redundancy of candidate sentences, where  $length(T)$

is the average length of all translation sentences and  $length(R)$  is the average length of all reference sentences.

$$Penalty = \begin{cases} \frac{length(R)}{length(T)} & \text{if } length(T) > length(R) \\ 1 & \text{Otherwise} \end{cases} \quad (4)$$

Then, the final score of  $C$  will be:

$$Score(C) = Re(C) \cdot Penalty \quad (5)$$

## 6 Experiments on MT System Diagnoses

In this section, to demonstrate the ability of our approach in the diagnoses of MT systems, we apply diagnostic evaluation to 3 statistical MT (SMT) systems and a rule-based MT (RMT) system respectively. We compare two SMT systems to understand the strength and shortcoming of each of them, and also compare a SMT system with the RMT system. The test corpus is NIST05 test data with 54852 check-points.

First SMT system (system A) is an implementation of classical phrase based SMT. The second SMT system (system B) shares the same decoder with system A and introduces a preprocess to reorder the long phrases in source sentences according to the syntax structure before decoding (Chiho Li et al., 2007). The third SMT system (system C) is a popular internet service and the RMT system (system D) is a popular commercial system.

In the first experiment, we diagnose the system A and B and compare the results as shown in table 7. When evaluated using BLEU, system B achieved a 0.005 points increase on top of system A which is not a very significant difference. The diagnostic results in table 7 provide much richer information. The results indicate that two systems perform similar at the word level categories while at all phrase level categories, system B performs better. This result reflects the benefit from the reordering of complex phrases in system B. Paired t-statistic score for each pair of category scores is also calculated by repeating the evaluation on a random copy of the test set with replacement (Koehn 2004). An absolute score beyond 2.17 of paired t-statistic means the difference of the samples is statistically significant (above 95%). Table 8 and 9 show an instance of the check-point and its evaluation in this experiment.

	System A	System B	T score
<b>WORDS</b>			
<b>Idiom</b>	0.1933	0.2370	13.38
<b>Adjective</b>	0.5836	0.5577	-17.43
<b>Pronoun</b>	0.7566	0.7344	-13.49
<b>Adverb</b>	0.5365	0.5433	7.11
<b>Preposition</b>	0.6529	0.6456	-6.21
<b>Repetitive word</b>	0.3363	0.3958	9.86
<b>PHRASEs</b>			
<b>Subject-predicate</b>	0.5117	0.5206	7.36
<b>Predicate-object</b>	0.4041	0.4180	15.52
<b>Predicate-complement</b>	0.4409	0.5125	9.51
<b>Measure phrase</b>	0.5030	0.5092	3.56
<b>Location phrase</b>	0.5245	0.5338	2.83
<b>GROUPs</b>			
<b>WORDS</b>	0.4839	0.4855	2.03
<b>PHRASEs</b>	0.4744	0.4964	13.97
<b>SYSTEM (Linguistic)</b>	0.4263	0.4370	16.50
<b>SYSTEM (BLEU)</b>	0.3564	0.3614	7.91

Table 7: Diagnose of SMT systems

Source Sentence	不过泰国总理戴克辛誓言将继续在其国内进行搜寻。
Category	Preposition_Object_Phrase
Check-Point	在其国内
Reference 1	in this country $DM = 0.5$
Reference 2	in his country $DM = 0.5$
System A Translation	but the prime minister of thailand Dex-in vowed to continue in domestic the search.
System B Translation	but the prime minister of thailand Dex-in vowed to continue the search in his country.

Table 8: An instance of the check-point.

	System A	System B
Ref 1: Match/Total	1/6	2/6
Ref 2: Match/Total	1/6	6/6
Score	0.17	1

Table 9: N-gram matching rate and scores.

Type	Normal	Lower	Stem
<b>Ambiguous word</b>	0.49/0.42	0.50/0.42	0.53/0.46
<b>New word</b>	0.13/0.13	0.37/0.32	0.42/0.35
<b>Idiom</b>	0.43/0.66	0.46/0.67	0.51/0.71
<b>Pronoun</b>	0.60/0.68	0.69/0.75	0.66/0.75
<b>Preposition</b>	0.38/0.42	0.42/0.45	0.43/0.46
<b>Collocation</b>	0.66/0.54	0.66/0.55	0.70/0.56
<b>Subject-predicate phrase</b>	0.46/0.30	0.51/0.36	0.58/0.42
<b>Predicate-object phrase</b>	0.37/0.25	0.37/0.26	0.47/0.29
<b>Compound sentence</b>	0.22/0.16	0.23/0.16	0.23/0.17

Table 10: Diagnose of SMT and RMT.

In the second experiment, we diagnose system C and D and compare the results. The BLEU score of system C is 0.3005 and system D is 0.2606. Table 10 shows the diagnostic results on categories with significant differences. Scores calculated with 3 matching options described in section 5 are given (“Lower” means Lowercase. The scores are listed in the form “SMT score/RMT score”). The diagnostic results indi-

cate that system C performs better on most categories than system D, but system D performs better on categories like idiom, pronoun and preposition. This result reveals a key difference between two types of MT systems: the SMT works well on the open categories that can be handled by context, while the RMT works well on closed categories which are easily translated by linguistic rules.

As the results of two experiments demonstrate, the diagnostic evaluation provides rich information of the capability of translating various important linguistic categories beyond a single system score. It successfully distinguishes the specific difference between the MT systems whose system level performance is similar. It can also diagnose the MT system with different architectures. Diagnostic evaluation tells the developers about the direction to improve the system. Along with the scores of categories, the diagnostic evaluation provides the system translation and references at every check-point so that the developers can trace and understand about how the MT system works on every single instance.

## 7 Experiments on Ranking MT Systems

Offering a general evaluation at system level is the major goal of state-of-the-art evaluation methods including widely accepted n-gram metrics. The absence of linguistic knowledge in BLEU motivated many work to integrate linguistic features into evaluation metric. In (Yang 2007), the evaluation of SMT systems is alternately formulated as a ranking problem. Different linguistic features are combined with BLEU such as matching rate of dependency relations of translation candidates against the reference sentences. The experiments demonstrate that the dependency matching rate feature can increase the ranking accuracy in some cases. Compared to dependency structure, the linguistic categories in our approach showcase more extensive features. It would be interesting to see whether the linguistic categories can be used to further improve the ranking of SMT systems.

In experiments, we use the scores of linguistic categories, dependency matching rate, scores of BLEU and other popular metrics as ranking features of MT systems and trained by Ranking SVM of SVMlight (Joachims, 1998). We performed the ranking experiments on ACL 2005 workshop data, ranking 7 MT translations with three-fold cross-validation both on sentence level and document level. The Spearman score is used

to calculate the correlation with human assessments. Table 11 and 12 show the results of the different feature sets on sentence level and document level respectively.

As shown in experiment results linguistic categories (LC), when used alone, are better related with human assessments than BLEU and GTM. When combined with the baseline metrics (BLEU & NIST), LC scores further improve the correlation score, better than dependence matching rate (DP). LC scores are obtained by matching the exact form of the words as METEOR(exact) does. NIST+LC combination score is better than METEOR(exact) at sentence and document level, and also better than METEOR(exact&syn) (syn means wn\_synonymy module in METEOR) at document level. This results indicate the ability of linguistic features in improving the performance of ranking task.

	Mean Correlation
BLEU 4	0.245
NIST 5	0.307
GTM (e=2)	0.251
METEOR(exact)	0.306
METEOR(exact&syn)	0.327
DP	0.246
LC	0.263
BLEU+DP	0.270
BLEU+ LC	0.288
BLEU+ DP +LC	0.307
NIST+ LC	0.322
NIST+ DP +LC	0.333

Table 11: Sentence level ranking (DP means dependency and LC means linguistic categories)

	Mean Correlation
BLEU 4	0.305
NIST 5	0.373
GTM (e=2)	0.327
METEOR(exact)	0.363
METEOR(exact&syn)	0.394
DP	0.323
LC	0.369
BLEU+DP	0.325
BLEU+ LC	0.387
BLEU+ DP +LC	0.332
NIST+ LC	0.409
NIST+ DP +LC	0.359

Table 12: Document level ranking

## 8 Comparison with Related Work

This work is inspired by (Yu, 1993) with many extensions. (Yu, 1993) proposed MTE evaluation system based on check-points for English-Chinese machine translation systems with human craft linguistic taxonomy including 3,200 pairs of sentences containing 6 classes of check-points. Their check-points were manually constructed by human experts, therefore it will be costly to build

new test corpus while the check-points in our approach are constructed automatically. Another limitation of their work is that only binary score is used for credits while we use n-gram matching rate which provides a broader coverage of different levels of matching.

There are many recent work motivated by n-gram based approach. (Callison-Burch et al., 2006) criticized the inadequate accuracy of evaluation at the sentence level. (Lin and Och, 2004) used longest common subsequence and skip-bigram statistics. (Banerjee and Lavie, 2005) calculated the scores by matching the unigrams on the surface forms, stemmed forms and senses. (Liu et al., 2005) used syntactic features and unlabeled head-modifier dependencies to evaluate MT quality, outperforming BLEU on sentence level correlations with human judgment. (Gimenez and Marquez, 2007) showed that linguistic features at more abstract levels such as dependency relation may provide more reliable system rankings. (Yang et al., 2007) formulates MT evaluation as a ranking problems leading to greater correlation with human assessment at the sentence level.

There are many differences between these n-gram based methods and our approach. In n-gram approach, a sentence is viewed as a collection of n-grams with different length without differentiating the specific linguistic phenomena. In our approach, a sentence is viewed as a collection of check-points with different types and depth, conforming to a clear linguistic taxonomy. Furthermore, in n-gram approach, only one general score at the system level is provided which make it not suitable for system diagnoses, while in our approach we can give scores of linguistic categories and provide much richer information to help developers to find the concrete strength and flaws of the system, in addition to the general score. The n-gram based metric is not very effective when comparing the systems with different architectures or systems with similar general score, while our approach is more effective in both cases by digging into the multiple linguistic levels and disclosing the latent differences of the systems.

## 9 Conclusion and Future Work

This paper presents an automatically diagnostic evaluation methods on MT based on linguistic check-points automatically constructed. In contrast with the metrics which only give a general score, our evaluation system can give developers

feedback about the faults and strength of an MT system regarding specific linguistic category or category group. Different with the existing work based on check-points, our work presents an approach to automatically generate the check-point database. We show that although there is some noise brought from word alignment and parsing, we can effectively alleviate the problem by refining the parser results, weighting the reference with confidence score and providing large quantity of check-points.

The experiments demonstrate that this method can uncover the specific difference between MT systems with similar architectures and different architectures. It is also demonstrated that the linguistic check-points can be used as new features to improve the ranking task of MT systems.

Although we present the diagnostic evaluation method with Chinese-English language pair, our approach can be applied to other language pair if syntax parser and word aligner are available.

The taxonomy used in current proposal is based on the human-made linguistic system. An interesting problem to be explored in the future is whether the taxonomy could be constructed automatically from the parsing results.

## References

- Statanjeev Banerjee, Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgements*. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization 2005.
- Chris Callison-Burch, Miles Osborne, Philipp Koehn. 2006. *Re-evaluating the Role of Bleu in Machine Translation Research*. In Proceedings of the European Chapter of the ACL 2006.
- Martin Chodorow, Claudia Leacock. 2000. *An unsupervised method for detecting grammatical errors*, In 1st Meeting of the North America Chapter of the ACL, pp.140–147, 2000.
- Thorsten Joachims. 1998. *Making Large-scale Support Vector Machine Learning Practical*, In B. Scholkopf, C. Burges, A. Smola. *Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, MA, December.
- Jesus Gimenez and Llis Marquez. 2007. *Linguistic features for automatic evaluation of heterogeneous MT systems*, Workshop of statistical machine translation in conjunction with 45<sup>th</sup> ACL, 2007.
- Dan Klein, Christopher Manning. 2003. *Accurate Unlexicalized Parsing*, Proceedings of the 41<sup>th</sup> Meeting of the ACL, pp. 423-430.
- Philipp Koehn. 2004. *Statistical Significance Tests for Machine Translation Evaluation*. In Proc. of the EMNLP, Barcelona, Spain.
- Chiho Li, Minghui Li, Dongdong Zhang, Mu Li, Ming Zhou, Yi Guan. 2007. *A Probabilistic Approach to Syntax-based Reordering for SMT*. In Proceedings of the 45<sup>th</sup> ACL, 2007.
- Chin-Yew Lin and Franz Josef Och. 2004. *Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics*. In Proceedings of the 42<sup>th</sup> ACL 2004.
- Ding Liu, Daniel Gildea. 2005. *Syntactic Features for Evaluation of Machine Translation*, ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.
- Shuxin Liu. 2002. *Linguistics of Contemporary Chinese Language (in Chinese)*, Advanced Education Publisher.
- Jiping Lv. 2000. *Foundation of Mandarin Grammar (in Chinese)*, Shangwu Publisher.
- Franz Josef Och, Hermann Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.
- Kishore Papieni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*, In Proceedings of the ACL 2002.
- Shiwen Yu. 1993. *Automatic evaluation of output quality for machine translation systems*, In Proceedings of the evaluators' forum, April 21-24, 1991, Les Rasses, Vaud, 1993.
- Yang Ye, Ming Zhou, Chinyew Lin. 2007. *Sentence level machine translation evaluation as a ranking problem: one step aside from BLEU*, In Workshop of statistical machine translation, in conjunction with 45<sup>th</sup> ACL, 2007.
- Ming Zhou. 2000, *A Block-Based Robust Dependency Parser for Unrestricted Chinese Text*. Proceedings of Second Chinese Language Processing Workshop, 2000, held in conjunction with ACL, 2000.