METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages

Michael Denkowski and **Alon Lavie**
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15232, USA
{mdenkows,alavie}@cs.cmu.edu

Abstract

This paper describes our submission to the WMT10 Shared Evaluation Task and MetricsMATR10. We present a version of the METEOR-NEXT metric with paraphrase tables for five target languages. We describe the creation of these paraphrase tables and conduct a tuning experiment that demonstrates consistent improvement across all languages over baseline versions of the metric without paraphrase resources.

1 Introduction

Workshops such as WMT (Callison-Burch et al., 2009) and MetricsMATR (Przybocki et al., 2008) focus on the need for accurate automatic metrics for evaluating the quality of machine translation (MT) output. While these workshops evaluate metric performance on many target languages, most metrics are limited to English due to the relative lack of lexical resources for other languages.

This paper describes a language-independent method for adding paraphrase support to the METEOR-NEXT metric for all WMT10 target languages. Taking advantage of the large parallel corpora released for the translation tasks often accompanying evaluation tasks, we automatically construct paraphrase tables using the *pivot* method (Bannard and Callison-Burch, 2005). We use the WMT09 human evaluation data to tune versions of METEOR-NEXT with and without paraphrases and report significantly better performance for versions with paraphrase support.

2 The METEOR-NEXT Metric

The METEOR-NEXT metric (Denkowski and Lavie, 2010) evaluates a machine translation hypothesis against a reference translation by calculating a similarity score based on an alignment between the two strings. When multiple references are provided, the hypothesis is scored against each and the reference producing the highest score is used. Alignments are formed in two stages: search space construction and alignment selection.

For a single hypothesis-reference pair, the space of possible alignments is constructed by identifying all possible word and phrase matches between the strings according to the following matchers:

Exact: Words are matched if and only if their surface forms are identical.

Stem: Words are stemmed using a language-appropriate Snowball Stemmer (Porter, 2001) and matched if the stems are identical.

Synonym: Words are matched if they are both members of a synonym set according to the WordNet (Miller and Fellbaum, 2007) database.

Paraphrase: Phrases are matched if they are listed as paraphrases in a paraphrase table. The tables used are described in Section 3.

Previously, full support has been limited to English, with French, German, and Spanish having exact and stem match support only, and Czech having exact match support only.

Although the exact, stem, and synonym matchers identify *word* matches while the paraphrase matcher identifies *phrase* matches, all matches can be generalized to phrase matches with a start position and phrase length in each string. A word occurring less than *length* positions after a match start is considered *covered* by the match. Exact, stem, and synonym matches always cover one word in each string.

Once the search space is constructed, the final alignment is identified as the largest possible subset of all matches meeting the following criteria in order of importance:

1. Each word in each sentence is covered by zero or one matches
2. Largest number of covered words across both

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 339–342,
Uppsala, Sweden, 15-16 July 2010.

sentences

3. Smallest number of chunks, where a chunk is defined as a series of matched phrases that is contiguous and identically ordered in both sentences

4. Smallest sum of absolute distances between match start positions in the two sentences (prefer to align words and phrases that occur at similar positions in both sentences)

Once an alignment is selected, the METEOR-NEXT score is calculated as follows. The number of words in the translation hypothesis (t) and reference (r) are counted. For each of the matchers (m_i), count the number of words covered by matches of this type in the hypothesis ($m_i(t)$) and reference ($m_i(r)$) and apply matcher weight (w_i). The weighted Precision and Recall are then calculated:

$$P = \frac{\sum_i w_i \cdot m_i(t)}{|t|} \qquad R = \frac{\sum_i w_i \cdot m_i(r)}{|r|}$$

The parameterized harmonic mean of P and R (van Rijsbergen, 1979) is then calculated:

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

To account for gaps and differences in word order, a fragmentation penalty (Lavie and Agarwal, 2007) is calculated using the total number of matched words (m) and number of chunks (ch):

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

The final METEOR-NEXT score is then calculated:

$$Score = (1 - Pen) \cdot F_{mean}$$

The parameters α, β, γ, and $w_i ... w_n$ can be tuned to maximize correlation with various types of human judgments.

3 The METEOR Paraphrase Tables

To extend support for WMT10 target languages, we use released parallel corpora to construct paraphrase tables for English, Czech, German, Spanish, and French. These tables are used by the METEOR-NEXT paraphrase matcher to identify additional phrase matches in each language.

3.1 Paraphrasing with Parallel Corpora

Following Bannard and Callison-Burch (2005), we extract paraphrases automatically from bilingual corpora using a *pivot phrase* method. For a given language pair, word alignment, phrase extraction, and phrase scoring are conducted on parallel corpora to build a single bilingual phrase table for the language pair. For each native phrase (n_1) in the table, we identify each foreign phrase (f) that translates n_1. Each alternate native phrase ($n_2 \neq n_1$) that translates f is considered a paraphrase of n_1 with probability $P(f|n_1) \cdot P(n_2|f)$. The total probability of n_2 paraphrasing n_1 is given as the sum over all f:

$$P(n_2|n_1) = \sum_f P(f|n_1) \cdot P(n_2|f)$$

The same method can be used to identify foreign paraphrases (f_1, f_2) given native pivot phrases n. To merge same-language paraphrases extracted from different parallel corpora, we take the mean of the corpus-specific paraphrase probabilities (P_C) weighted by the size of the corpora (C) used for paraphrase extraction:

$$P(n_2|n_1) = \frac{\sum_C |C| \cdot P_C(n_2|n_1)}{\sum_C |C|}$$

To improve paraphrase accuracy, we apply multiple filtering techniques during paraphrase extraction. The following are applied to each paraphrase *instance* (n_1, f, n_2):

1. Discard paraphrases with very low probability ($P(f|n_1) \cdot P(n_2|f) < 0.001$)

2. Discard paraphrases for which n_1, f, or n_2 contain *any* punctuation characters.

3. Discard paraphrases for which n_1, f, or n_2 contain *only* common words. Common words are defined as having relative frequency of 0.001 or greater in the parallel corpus.

Remaining phrase instances are summed to construct corpus-specific paraphrase tables. Same-language paraphrase tables are selectively merged as part of the tuning process described in Section 4.2. Final paraphrase tables are further filtered to include only paraphrases with probabilities above a final threshold (0.01).

Language Pair		Corpus	Phrase Table
Target	Source	Sentences	Phrase Pairs
English	Czech	7,321,950	128,326,269
English	German	1,630,132	84,035,599
English	Spanish	7,965,250	363,714,779
English	French	8,993,161	404,883,736
German	Spanish	1,305,650	70,992,157

Table 1: Sizes of training corpora and phrase tables used for paraphrase extraction

Language	Pivot Languages	Phrase Pairs
English	German, Spanish, French	6,236,236
Czech	English	756,113
German	English, Spanish	3,521,052
Spanish	English, German	6,352,690
French	English	3,382,847

Table 2: Sizes of final paraphrase tables

3.2 Available Data

We conduct paraphrase extraction using parallel corpora released for the WMT10 Shared Translation Task. This includes Europarl corpora (French-English, Spanish-English, and German-English), news commentary (French-English, Spanish-English, German-English, and Czech-English), United Nations corpora (French-English and Spanish-English), and the CzEng (Bojar and Žabokrtský, 2009) corpus sections 0-8 (Czech-English). In addition, we use the German-Spanish Europarl corpus released for WMT08 (Callison-Burch et al., 2008).

3.3 Paraphrase Table Construction

Using all available data for each language pair, we create bilingual phrase tables for the following: French-English, Spanish-English, German-English, Czech-English, and German-Spanish. The full training corpora and resulting phrase tables are described in Table 1. For each phrase table, both foreign and native paraphrases are extracted. Same-language paraphrases are selectively merged as described in Section 4.2 to produce the final paraphrase tables described in Table 2. To keep table size reasonable, we only extract paraphrases for phrases occurring in target corpora consisting of the pooled development data from the WMT08, WMT09, and WMT10 translation tasks (10,158 sentences for Czech, 20,258 sentences for all other languages).

Target	Systems	Usable Judgments
English	45	20,357
Czech	5	11,242
German	11	6,563
Spanish	9	3,249
French	12	2,967

Table 3: Human ranking judgment data from WMT09

4 Tuning METEOR-NEXT

4.1 Development Data

As part of the WMT10 Shared Evaluation Task, data from WMT09 (Callison-Burch et al., 2009), including system output, reference translations, and human judgments, is available for metric development. As metrics are evaluated primarily on their ability to rank system output on the segment level, we select the human ranking judgments from WMT09 as our development set (described in Table 3).

4.2 Tuning Procedure

Tuning a version of METEOR-NEXT consists of selecting parameters (α, β, γ, $w_i ... w_n$) that optimize an objective function for a given language. If multiple paraphrase tables exist for a language, tuning also requires selecting the optimal set of tables to merge.

For WMT10, we tune to rank consistency on the WMT09 data. Following Callison-Burch et. al (2009), we discard judgments where system outputs are deemed equivalent and calculate the proportion of remaining judgments preserved when system outputs are ranked by automatic metric scores. For each target language, tuning is conducted as an exhaustive grid search over metric parameters and possible paraphrase tables, resulting in global optima for both.

5 Experiments

To evaluate the impact of our paraphrase tables on metric performance, we tune versions of METEOR-NEXT with and without the paraphrase matchers for each language. For further comparison, we tune a version of METEOR-NEXT using the TERp English paraphrase table (Snover et al., 2009) used by previous versions of the metric.

As shown in Table 4, the addition of paraphrases leads to a better tuning point for every target language. The best scoring subset of paraphrase ta-

Language	Paraphrases	Rank Consistency	α	β	γ	w_{exact}	w_{stem}	w_{syn}	w_{par}
English	none	0.619	0.85	2.35	0.45	1.00	0.80	0.60	–
	TERp	0.625	0.70	1.40	0.25	1.00	0.80	0.80	0.60
	de+es+fr	**0.629**	0.75	0.60	0.35	1.00	0.80	0.80	0.60
Czech	none	0.564	0.95	0.20	0.70	1.00	–	–	–
	en	**0.574**	0.95	2.15	0.35	1.00	–	–	0.40
German	none	0.550	0.20	0.75	0.25	1.00	0.80	–	–
	en+es	**0.576**	0.75	0.80	0.90	1.00	0.20	–	0.80
Spanish	none	0.586	0.95	0.55	0.90	1.00	0.80	–	–
	en+de	**0.608**	0.15	0.25	0.75	1.00	0.80	–	0.40
French	none	0.696	0.95	0.80	0.35	1.00	0.60	–	–
	en	**0.707**	0.90	0.85	0.45	1.00	0.00	–	0.60

Table 4: Optimal METEOR-NEXT parameters with and without paraphrases for WMT10 target languages

bles for English also outperforms the TERp paraphrase table.

Analysis of the phrase matches contributed by the paraphrase matchers reveals an interesting point about the task of paraphrasing for MT evaluation. Despite filtering techniques, the final paraphrase tables include some unusual, inaccurate, or highly context-dependent paraphrases. However, the vast majority of matches identified between actual system output and reference translations correspond to valid paraphrases. In many cases, the evaluation task itself acts as a final filter; to produce a phrase that can match a spurious paraphrase, not only must a MT system produce incorrect output, but it must produce output that overlaps exactly with an obscure paraphrase of some phrase in the reference translation. As systems are far more likely to produce phrases with similar words to those in reference translations, far more valid paraphrases exist in typical system output.

6 Conclusions

We have presented versions of METEOR-NEXT and paraphrase tables for five target languages. Tuning experiments indicate consistent improvements across all languages over baseline versions of the metric. Created for MT evaluation, the METEOR paraphrase tables can also be used for other tasks in MT and natural language processing. Further, the techniques used to build the paraphrase tables are language-independent and can be used to improve evaluation support for other target languages. METEOR-NEXT, the METEOR paraphrase tables, and the software used to generate paraphrases are released under an open source license and made available via the METEOR website.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proc. of ACL05*.

Ondřej Bojar and Zdeněk Žabokrtský. 2009. CzEng 0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc. of WMT08*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of WMT09. In *Proc. of WMT09*.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR Machine Translation Metric to the Phrase Level for Improved Correlation with Human Post-Editing Judgments. In *Proc. NAACL/HLT 2010*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proc. of WMT07*.

George Miller and Christiane Fellbaum. 2007. WordNet. http://wordnet.princeton.edu/.

Martin Porter. 2001. Snowball: A language for stemming algorithms. http://snowball.tartarus.org/texts/.

M. Przybocki, K. Peterson, and S Bronsart. 2008. Official results of the NIST 2008 "Metrics for MAchine TRanslation" Challenge (MetricsMATR08).

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. In *Proc. of WMT09*.

C. van Rijsbergen, 1979. *Information Retrieval*, chapter 7. 2nd edition.