BLEU+
An Extension of BLEU for Agglutinative Languages

Introduction

BLEU (Papineni, Roukos et al. 2002) is widely used in Machine Translation (MT) community because it is an automatic method to evaluate the performance of the translation and it is reported to be in correlation with human judges.

Recently, some negative aspects of BLEU method is published (Banerjee and Lavie 2005; Callison-Burch, Osborne et al. 2006), but it is still the most appropriate and common way to report the efficiency of an MT system.

In this software project, we have implemented a modified version of BLEU for agglutinative languages. in order to built a bit more "intelligent" way to find "word matches". In agglutinative languages like Turkish, words are combination of a root and a variable number suffixes affixed to that root. The straight-forward matching technique in BLEU fails when one suffix is misplaced or mistranslated, even if the roots and all other suffixes are matching. Our BLEU+ evaluation tool accepts words in decomposed form (root+suffix1+suffix2+...) and applies a more intelligent strategy for matching by using some fine-grained matching procedure.

Agglutinative Languages

In agglutinative languages, suffixes are affixed to a root word or another suffix sequentially, just like “beads on a string”. These suffixes can have modifies the root in some aspects or adds additional information to the root. Languages like Turkish, Finnish, Hungarian and Estonian is classified as agglutinative languages.

Because of the relatively complex morphology, the number of surface forms of the words (or wordforms) are much much larger than English, German or any other language. For instance, since Turkish has a very productive derivational and inflectional morphology, the number of surface forms that can be generated by using only one root word is more than one million (Hankamer 1986). Here is a daily example of Turkish word :

evlerimizdekilerden (from the ones in our houses)

ev+Noun+A3pl+P1pl+Loc^DB+Adj+Rel^DB+Noun+Zero+A3pl+Pnon+Abl

Classical BLEU Evaluation

In its basic form, BLEU evaluation is based on a clipped precision metric which is calculated by counting the matching words in candidate sentence and reference sentence(s). This matching procedure is a straight-forward match which imposes a full surface form match. However, this harsh mathcing usually causes sharp deficiencies in BLEU scores, even though the actual quality of the candidate sentence is not that bad.

BLEU+ Strategy

In order to alleviate this problem, we developed a more "intelligent" way to count the matching words. The main idea behind the method is taking the un-matched surface forms into account by applying some levels of penalty multipliers. Certainly, it is required to have a matching root word. If a candidate root word does not occur in any of the references, it is a total unmatch. Our method tries to count the words with small suffix differences between their reference equivalents. For example :

Candidate : ev+ler+de ……. (At the houses …..)
Reference : ev+de …….. (At the house ……)

In the example above, there is not big semantic shift in the candidate, but this candidate wordform can not contribute to total BLEU score, because there is not a match on surface form basis.

BLEU+ tries to handle this situations by using this method:

1 Try to find a surface match in any of the references

2 If a surface match (full match) can not be found, try find a word with the same root word in any of the references.

2.1. If such a reference word is found, calculate the distance ***d*** (suffix based Levenshtein distance) between them

2.1.1. If ***d*** is below a threshold t, count this as "a partial match" (multiplied by a penalty p_d, where $0 \leq p_d \leq 1$).

2.1.2. If ***d*** is above threhold ***t***, count this as a "only root word match" (multiplied by a penalty p_{root}, where $0 \leq p_{root} \leq 1$)

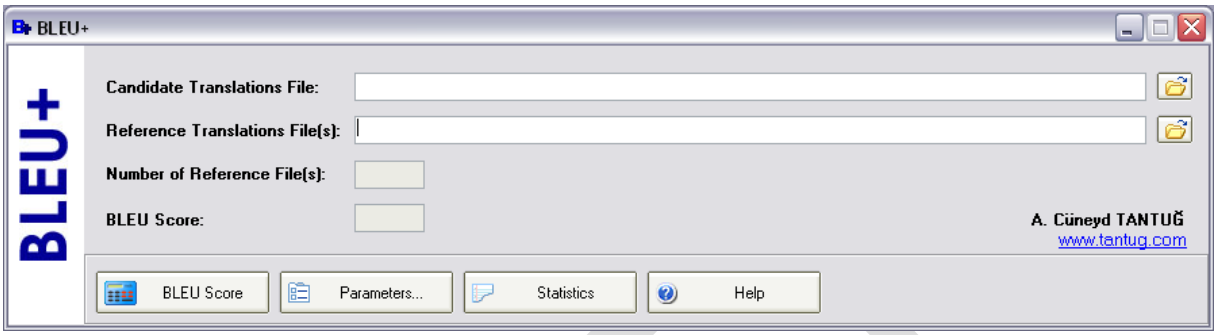
The parameters of our algorithm are :

t **:** threshold value for the maximum suffix based distance between words
p_1 **:** penalty multiplier for partial matches with d=1
p_2 **:** penalty multiplier for partial matches with d=2
…
p_d **:** penalty multiplier for partial matches with d
p_{root} **:** penalty multiplier for only root matches (d > t)

The distance between n-gram units with n>1 is the sum of distances between all individual unigrams composing the whole n-gram structure.

BLEU+ Tool

The BLEU+ method is implemented in a BLEU+ Tool together with a standart BLEU implementation. This tool enables for multiple reference files as well as one reference file.

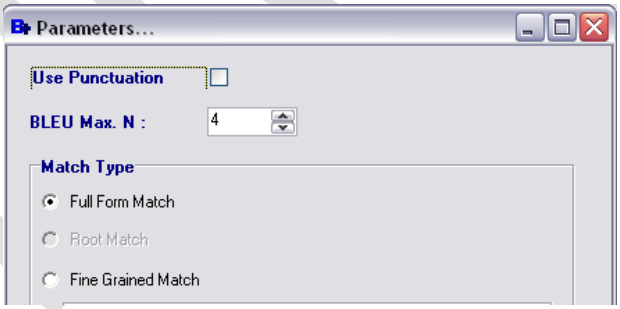


Candidate Translations File : The file containing the outputs of the system must be specified in this field

Referemce Translations File(s) : The file(s) containing the reference translations must be specified by using the button on the right of the field. User can select multiple files in the pop-up file selection window.

BLEU Score : Calculate the BLEU score with the specified set of parameters.

Parameters : User can change parameters by using this button. The parameter is screenshot of parameter screen is given below :



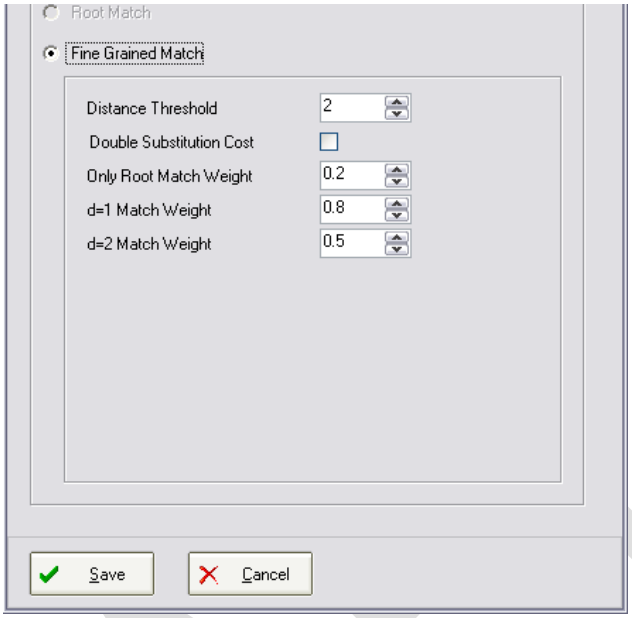
Use Punctuation : In order to take the punctuation into account, this check box should be clicked.

BLEU Max. N : The number of desired n-gram size can be set by using this field. The default value is 4 (as described in the (Papineni, Roukos et al. 2002)). Maximum value is 5.

Match Type : User can specify the matching strategy.

Full Form Match : Counts only surface forms matches (calculates original BLEU score)

Fine-Grained Match : Implements BLEU+ matching strategy.



Distance Threshold : t parameter**.** Must be between 0 and 8.

Double Substitiution Cost : If this check box ise checked, the substitution operation is considered as a result of two sequential operations (delete and add) and this doubles the cost of substitution. Otherwise, substitution operation is a regular operation in calculating Levenshtein distance.

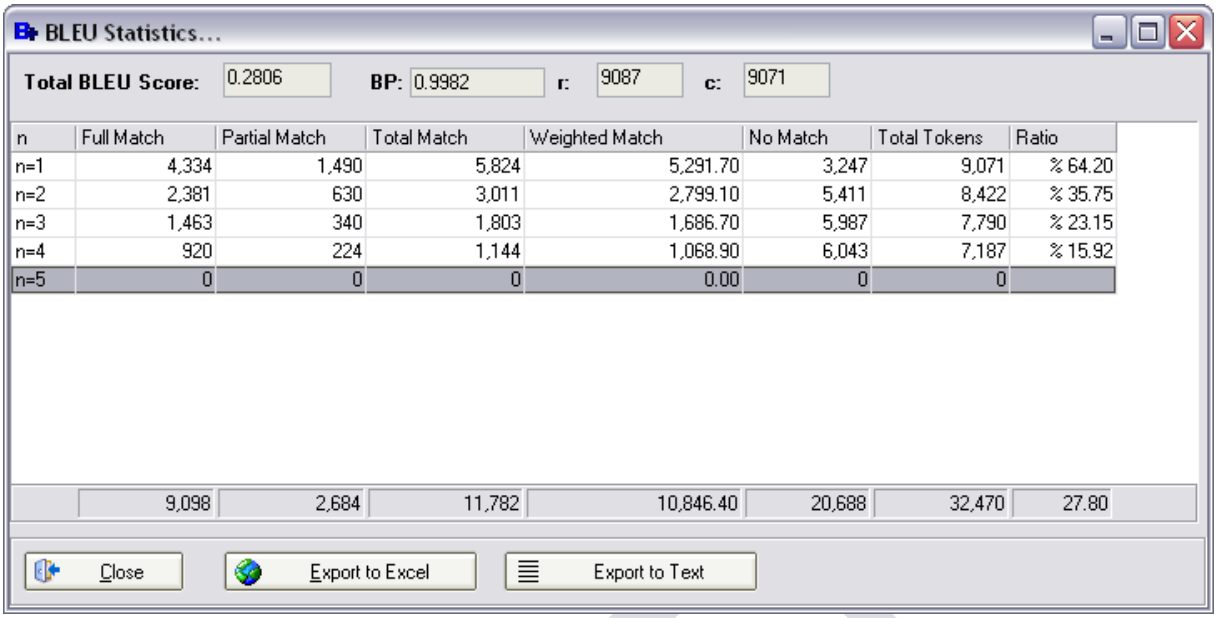
Only Root Match Weight : This is the penalty value used for words that has a root match in reference but the distance is above threshold. Value must between 0 and 1.

d=1 Match Weight : This is the penalty value for the candidate words that have only one different suffix than the corresponding reference word. Value must between 0 and 1.

d=2 Match Weight : This is the penalty value for the candidate words that have only one different suffix than the corresponding reference word. Value must between 0 and 1.

Up to 8 distance weights can be specified.

Statistics : After calculation of BLEU or BLEU+ score, the user can see the details of the calculation by using this button.



BP : Brevity Penalty

r: Effective reference length (calculated over entire reference senteces set)

c: Candidate length (calculated over entire reference candidate set)

Full Match : Number of clipped counts of surface form matches

Partial Match : Number of clipped counts of partial matches (with $d \leq t$)

Total Match : Total number of clipped matches (full match+partial match)

Weighted Match : Weighted number of clipped matches
(full match+ p_d x partial match)

No Match : Number of unmatched tokens in candidate sentences

Total Tokens: Number of tokens in candidate sentences

Ratio: Percentage of weighted match and total
(100 x Weighted Match / Total Tokens)

Licence

The current verison of BLEU+ tool (ver. 1.1) is distributed with a freeware licence.

Further Questions

Please contact tantug@itu.edu.tr for your further questions.

References

Banerjee, S. and A. Lavie (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, MI, USA.

Callison-Burch, C., M. Osborne, et al. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. Conference of the European Chapter of the Association for Computational Linguistics (EACL'06). Trento, Italy.

Hankamer, J. (1986). Finite State Morphology and Left to Right Phonology. West Coast Conference on Formal Linguistics Forum, Stanford University.

Papineni, K., S. Roukos, et al. (2002). BLEU : A Mehtod for Automatic Evaluation of Machine Translation. Association of Computational Linguistics, ACL'02. Philadelphia, PA, USA.