BLEU+: A Tool for Fine-Grained BLEU Computation

A. Cüneyd Tantuğ, Kemal Oflazer, İlknur D. El-Kahlout

Istanbul Technical University, Sabancı University, Sabancı University
Istanbul, Turkey
tantug@itu.edu.tr, oflazer@sabanciuniv.edu, ilknurdurgar@su.sabanciuniv.edu

Abstract

We present a tool, BLEU+, which implements various extension to BLEU computation to allow for a better understanding of the translation performance, especially for morphologically complex languages. BLEU+ takes into account both "closeness" in morphological structure, "closeness" of the root words in the WordNet hierarchy while comparing tokens in the candidate and reference sentence. In addition to gauging performance at a finer level of granularity, BLEU+ also allows the computation of various upper bound oracle scores: comparing all tokens considering only the roots allows us to get an upper bound when all errors due to morphological structure are fixed, while comparing tokens in an error-tolerant way considering minor morpheme edit operations, allows us to get a (more realistic) upper bound when tokens that differ in morpheme insertions/deletions and substitutions are fixed. We use BLEU+ in the fine-grained evaluation of the output of our English-to-Turkish statistical MT system.

1. Introduction

Evaluation is one of the most challenging problems in machine translation. The best way of evaluating an MT system is ultimately based on human judgement with which aspects of translation quality such as adequacy, fidelity and fluency can be judged. Human evaluation is however slow and labor intensive. BLEU (Papineni et al., 2002) has been proposed and used as an automatic way of gauging MT quality.[1] BLEU scores output of an MT system by comparing each sentence to a set reference translations using n-gram overlaps of word sequences. If a lot of words in the candidate occur in the reference, then the candidate is considered "adequate" while if a lot of n-grams of words (especially for "large" n) occur in the reference, then the candidate is considered "fluent".

While word-to-word comparisons in computing such overlaps are meaningful for some language pairs, the all-or-none nature of word comparisons can be particularly harsh for a morphologically complex target language, when the translation system generates sequences of morphemes that make up target words. BLEU – comparing the words – can flag a word as a mismatch although for example the corresponding target and reference morpheme sequences may contain morphemes with very close morphosemantics. Considering such cases as a complete mismatch also downgrades the performance of the system even though it gets most of the morphemes correctly.

In this paper we present a tool, BLEU+, which implements various extension to BLEU computation to allow for a better understanding of the translation performance, especially for morphologically complex languages.[2] BLEU+ takes into account both "closeness" in morphological structure, "closeness" of the root words in the WordNet hierarchy while comparing tokens in the candidate and reference sentence. In addition to gauging performance at a finer level of granularity, BLEU+ also allows the computation of various upper bound oracle scores: comparing all tokens considering only the roots allows us to get an upper bound when all errors due to morphological structure are fixed, while comparing tokens in an error-tolerant way considering minor morpheme edit operations, allows us to get a (lower) upper bound when tokens that differ in morpheme insertions/deletions and substitutions are fixed. We use BLEU+ in the fine-grained evaluation of the output of our English-to-Turkish statistical MT system (Oflazer and Durgar El-Kahlout, 2007).

1.1. Turkish

Turkish is an Ural-Altaic language, having highly agglutinative word structures with productive inflectional and derivational processes. It may be possible to have a single word translation of a sentence in English; e.g. *sağlamlaştırabiliyorlardı* (*sağlam+laş+tır+abil+iyor+lar+dı* split into morphemes) would be a translation of `They were able to strengthen it.` Our English-to-Turkish statistical MT system (Oflazer and Durgar El-Kahlout, 2007) produces a translation as a sequence of root and suffixes and each root is then concatenated with the following suffixes in the given order to produce words. One can appreciate that if all but one morpheme in the (synthetic) example above was incorrect, the whole word would be incorrect though the MT system clearly is doing some things correctly. For example the word *sağlam+laş+tır+abil**+mekte**+lar+dı.* has essentially the same meaning but would be counted as incorrect. Note that a morpheme level BLEU score could be used but that may be overly optimistic and not necessarily be comparable to word-level BLEU.

The following candidate and reference translations (with both word-level (W) and lexical-morpheme (L) representations) exemplify the problem more acutely.

[1] Though BLEU has certainly many problems as described by Callison-Burch et al. (Callison-Burch et al., 2006).

[2] This tool can be downloaded from `http://ddi.ce.itu.edu.tr`.

Candidate	(W) iki aile arasındaki **husumet** ve **kavga** uzun **yıllardır** sürü**yor**du. (L)iki aile ara+sh+nda+ki husumet ve kavga uzun **yıl+lar+dhr** sür**+hyor**+dh.
Reference	(W)iki aile arasında düşmanlık ve çatışma uzun **senelerdir** sür**mekte**ydi. (L) iki aile ara+sh+nda düşmanlık ve çatışma uzun **sene+lar+dhr** sür**+makta**+ydh. (The hostility and fight between two families had been lasting for many years.)

In the candidate translation, we have 4 of the last 6 words not matching the corresponding word in the reference translation. However, *husumet* (enmity) is a synonym of the reference word *düşmanlık* while *kavga* (fight) is a hyponym of *çatışma* (confrontation) in the Turkish WordNET (Bilgin et al., 2004). Also, the roots *yıl* (year) and *sene* are synonyms in the inflected words *yıllardır* (for years) and *senelerdir*.[3] Finally, the verb of the sentence in the candidate and the reference look different, but the difference is due to the use of the two almost synonymous morphemes. For all practical purposes, the candidate translation sentence renders the same meaning as the reference sentence but BLEU would consider as having a significant mismatch.

2. BLEU+

In order to alleviate the shortcoming of strict word-based matching used by the standard BLEU measure for languages like Turkish, we formulated and implemented an extension, BLEU+, that can perform finer-grained lexical comparison taking into synonymous roots (as in METEOR (Banerjee and Lavie, 2005)) and synonymous *morphemes*. Furthermore BLEU+ can take into consideration hypernym and hyponym relations between candidate and reference root words and compute various oracle scores. BLEU+ tool can also generate scores based on the METEOR metric (Banerjee and Lavie, 2005), a recent evaluation metric which incorporates the recall scores into evaluation process.

2.1. Matching Root Words

Assuming the candidate and reference translations are available in a morphemic representation, BLEU+ goes beyond the all-or-none nature of the word-to-word match in the computation of BLEU. Whenever a WordNET ontology is available, one can ask for root word matches based on synonymity. One can also consider hypernyms or hyponyms of a root word and these are not necessarily limited to the immediately neighboring root words but words that can be further away. The user may or may not choose to penalize synonymous matches and can define a detailed match penalties based on the distance of the matching word from the candidate word in the WordNET hypernym/hyponym hierarchy.

[3]Note also that the lexical morphemes also surface differently in these words, due to morhophonological processes such as vowel harmony, etc.

2.2. Matching Morphemes

BLEU+ can also identify matches of (almost) synonymous morphemes. For instance, in the example above, the lexical morphemes *+hyor* and *+makta* are almost synonymous, although the resulting surface forms are different. BLEU+ takes a list of such synonymous morphemes (along with possible contextual restrictions as to when they can be considered synonymous) and computes (penalized) matches taking into account such synonymous morphemes.

2.3. Scoring Matches

In the standard BLEU computation, words are compared with strict string equality comparison. In BLEU+, we define the *similarity* of a candidate word w_i with a reference word w_j, $S(w_i, w_j)$ as follows:

$$S(w_i, w_j) = S_{\text{root}}(w_i, w_j) \times S_{\text{morph}}(w_i, w_j)$$

Here $S_{\text{root}}(w_i, w_j)$ measures the similarity of the root words is based on a slightly more general notion: It evaluates to 1 if the *roots* of the words are the same but if the roots are synonyms, hypernyms or hyponyms, they are still considered similar but the score can be set to a value less than 1 depending on how much we want to discount synonymous and other matches. $S_{\text{root}}(w_i, w_j)$ would obviously be 0 if the root words are not related at all.
$S_{\text{morph}}(w_i, w_j)$ measures the similarity of the respective morphemes in the two words and is the product of the similarities of individual bound morphemes in both words. If two morphemes are the same, then their similarity is 1. If, however, the morphemes are not same but morphosemantically close as defined by a set of rules, then the similarity can be set to a value less than 1.

2.4. Computing Oracle Scores

Since translation into Turkish involves getting both the word sequence right and then getting the morpheme sequence in each word right, it may be helpful if we could factor out the contributions of the errors in each of these processes, as solutions to these individual problems require vastly different mechanisms.
If we compute BLEU scores only considering the roots, that would give us a high(er) oracle BLEU score which indicates the maximum score we would get if we got the morphemes and their order perfectly correct for each word. In this case, the similarity measure above is taken as

$$S(w_i, w_j) = S_{\text{root}}(w_i, w_j)$$

Another oracle score that we can compute is based on identifying words whose roots are similar as defined above but the morphological structure of the words are different. If the morpheme sequences differ by a small amount, that is, one morpheme sequence can be obtained from the other, by a small number of morpheme insertions, deletions or substitutions, then it may be worthwhile to identify some of these cases and attempt to correct them (akin to spelling correction but at the morpheme level). This oracle score gives us the maximum BLEU score that we can obtain if we can identify and "fix" all words whose roots are similar but the

morpheme structures differ by a small number of edit operations (usually 1 or 2). In fact, we have already benefitted from knowing this oracle score and have implemented various techniques to do "word-repair" (Oflazer, 2008).

2.5. Comparison Logs

BLEU+ also produces a fully annotated XML file that shows how the sentences match and indicates words that are assumed to match due to synonym, hyponym or hypernym relationships, morpheme synonymity or close synoymity and also indicates words whose morphology can be repaired to match the reference word. This file can then be mined in various ways for a detailed analysis of any errors. Figure 1 shows an example sentence annotation from a log file.

3. BLEU+ GUI

BLEU+ provides a graphical user interface through which various options can be set. Figure 2 shows the main screen of BLEU+ where the candidate and reference sentence files can be selected and the evaluation results can be seen.

Figure 3 shows the main parameter options for BLEU+. Here one can opt to consider or not to consider punctuation tokens in the BLEU computation, set the maximum number of n-grams to use, select how to perform token comparisons and also opt to computer METEOR scores. The *Surface Form Match* corresponds to the standard BLEU word comparison. *BLEU+ (Only Root Match)* will match only the roots of the tokens (and hence will compute the first oracle score above), while *BLEU+ (Fine Grained Match)* will compare at the morpheme level and use WordNet information if such a database is available.

Figure 4 shows the parameter settings for root matching using a WordNet database. Here one can select the WordNet file to use, and set any penalties for synonym/hypernym/hyponym matching.

Figure 5 shows the parameter settings for matching morphemes or combinations of morphemes. We provide one entry for each pair of morphemes or morpheme combinations that we consider to be semantically close and provide penalty figures for each pair.

Finally Figure 6 shows settings for oracle computations for morphological close word matches. Here we can select the morpheme edit distance to use when matching words whose roots match but morpheme sequences are close.

4. Using BLEU+

We used BLEU+ to evaluate in detail, the 649 sentence test set from English-Turkish SMT system. To exploit morphology, we used the morphemic representation of sentences in the evaluation. The basic BLEU score based on the standard definition of BLEU is 27.64 (54.90/31.90/21.9/15.90). Table 1 shows the BLEU+ scores when synonymous morpheme pairs and and WordNet relations included in the evaluation. We also provide oracle scores for root matching and 1 and 2 morpheme correction matches.

Table 1: BLEU+ scores

Matching	*BLEU+*	*N-gram precisions*
Default BLEU	27.64	54.90/31.90/21.9/15.90
Syn. Morphemes	27.79	55.20/32.00/22.00/16.00
WordNet	27.94	55.67/32.20/22.10/16.10
Combined	28.10	56.00/32.50/22.20/16.20
Root (oracle)	33.12	66.80/38.50/26.00/18.90
Morp. Corr. d=1 (ora.)	32.40	62.90/37.70/25.70/18.90
Morp. Corr. d=2 (ora.)	33.03	66.00/38.50/26.00/18.90

5. Conclusions

In this paper, we present a fine-grained extension to the industry standard BLEU MT evaluation metric, and a tool that implements it. This tool BLEU+ provides various fine-grained analyses of candidate translation by taking into account synonymous roots, and morphemes, and can compute oracle scores to show upper bound performance. It is especially helpful in providing a detailed understanding of translation performance into morphologically complex languages like Turkish, Hungarian, or Finnish especially when translation output is generated as a sequence of morphemes. Currently we are conducting some tests to show the correlation of our BLEU+ scores with human evaluation scores.

6. Acknowledgements

This work was in part supported by TÜBİTAK – The Turkish National Science and Technology Foundation under project grant 106E048 to Istanbul Technical University and project grant 105E020 to Sabancı University.

7. References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA, June. Association for Computational Linguistics.

Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer. 2004. Building a Wordnet for Turkish. *Romanian Journal of Information Science and Technology*, 7(1-2):163–172.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in Machine Translation research. In *EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento. Association for Computational Linguistics.

Kemal Oflazer and Ilknur Durgar El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.

Kemal Oflazer. 2008. Statistical machine translation into a morphologically complex language. In *Computational Linguistics and Intelligent Text Processing, 9th International Conference, CICLing 2008, Haifa, Israel*, volume 4919 of *Lecture Notes in Computer Science*, pages 376–387.

Kishore Papineni, Todd Ward Salim Roukos, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of

<SENTENCE id=”2”>
<FRAGMENTATION> 5 / 5 = 1,00 </FRAGMENTATION>
<CANDIDATE>
(radyo) / tv yayın+ch+lhk+sh+locative+accusative (yasakla+yan) bir (türk) vatandaş+sh taraf+sh+ablative kullan+hl+yan yasal hüküm+lar+nhn (kaldır+hl+ma+sh) anadil+dhr (.)
</CANDIDATE>
<REFERENCE1>
(türk) vatandaş+lar+sh+nhn televizyon ve (radyo) yayın+ch+lhk+sh+locative anadil+lar +sh+accusative kullan+ma+larh+accusative (yasakla+yan) hukuki düzenle+ma+lar var i +sa (kaldır+hl+ma+sh) (.)
</REFERENCE1>
<N n=”1”>
<FULLMATCH c="1"> radyo </FULLMATCH>
<NOMATCH> / </NOMATCH>
<NOMATCH> tv </NOMATCH>
<REPAIRED_MATCH d=”1” original=”yayın+ch+lhk+sh+locative+accusative” > yayın+ch+lhk+sh+locative </REPAIRED_MATCH>
<FULLMATCH c="1"> yasakla+yan </FULLMATCH>
<NOMATCH> bir </NOMATCH>
<FULLMATCH c="1"> türk </FULLMATCH>
<REPAIRED_MATCH d="2" original="vatandaş+sh"> vatandaş+lar+sh+nhn </REPAIRED_MATCH>
<NOMATCH> taraf+sh+ablative </NOMATCH>
<NOMATCH> kullan+hl+yan </NOMATCH>
<NOMATCH> yasal </NOMATCH>
<NOMATCH> hüküm+lar+nhn </NOMATCH>
<FULLMATCH c="1"> kaldır+hl+ma+sh </FULLMATCH>
<NOMATCH> anadil+dhr </NOMATCH>
<FULLMATCH c="1"> . </FULLMATCH>
</N>
</SENTENCE>

Figure 1: BLEU+ Log File

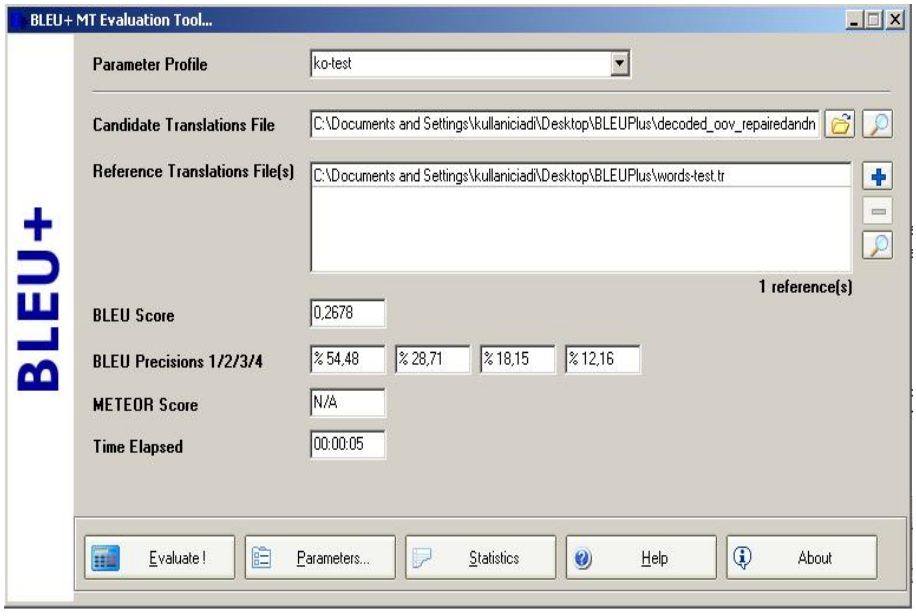

Figure 2: BLEU+ Evaluation Window

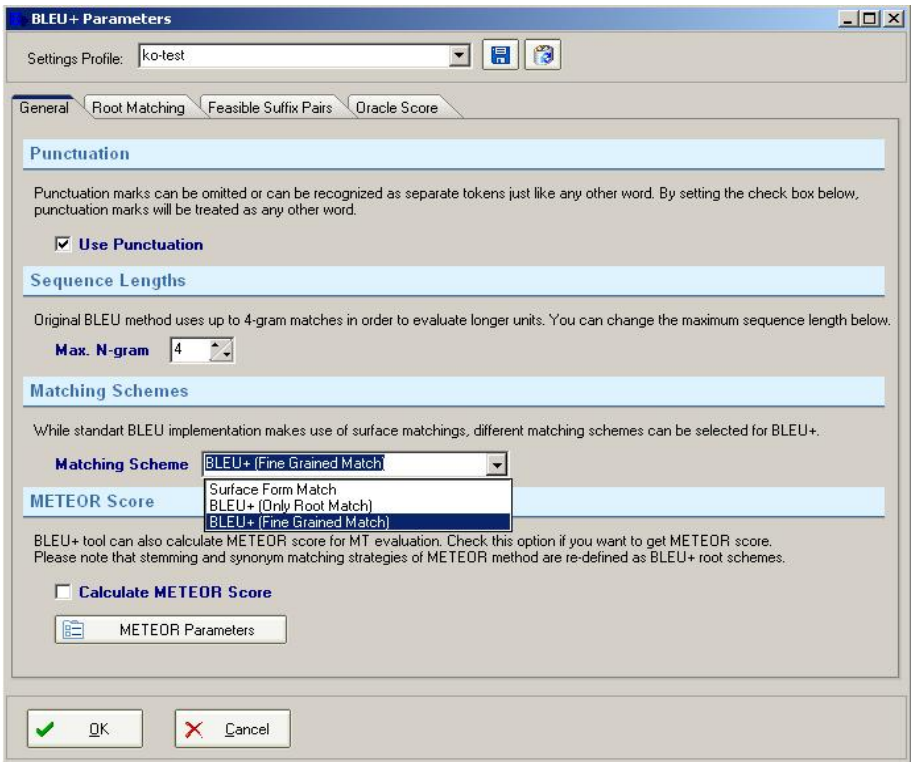

Figure 3: BLEU+ parameters

machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 311–318, Philadelphia, July. Association for Computational Linguistics.

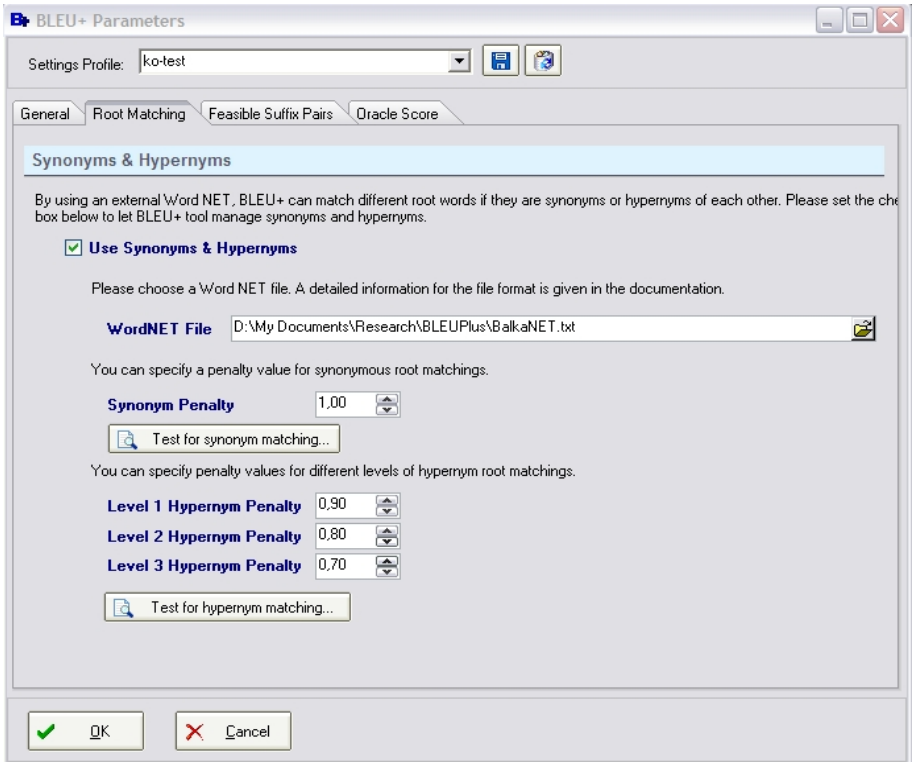

Figure 4: Root Matching Parameters

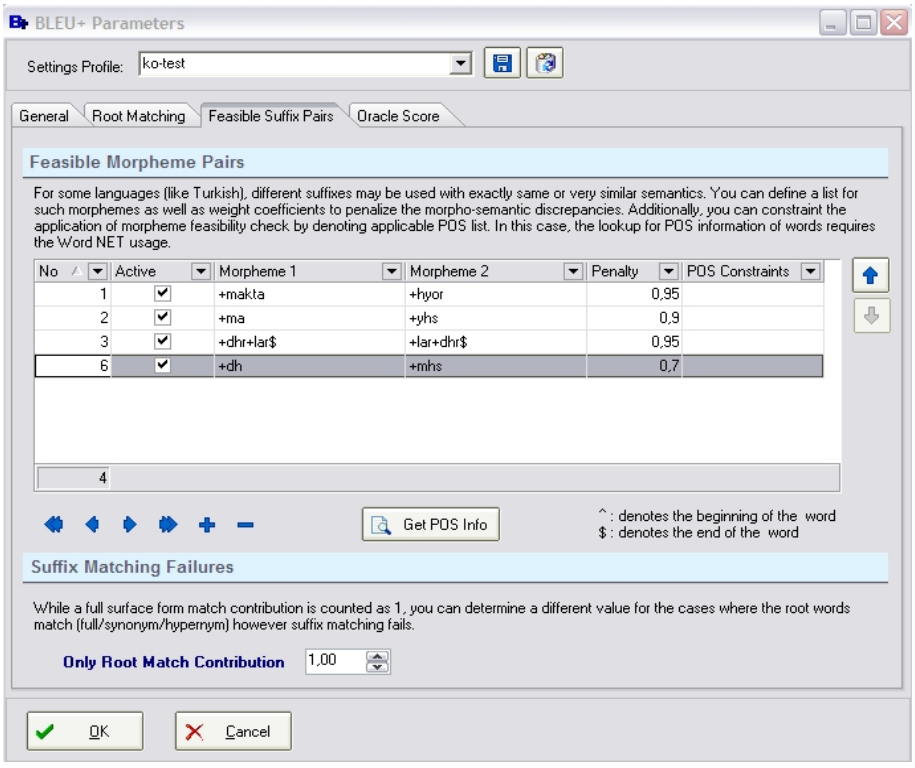

Figure 5: Morpheme Matching Parameters

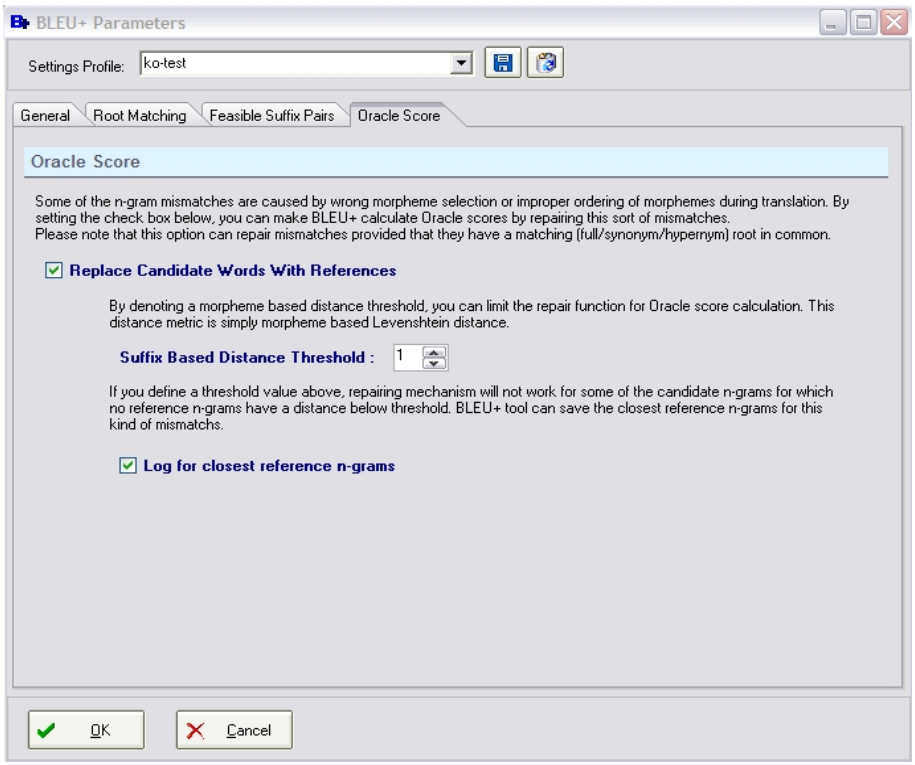

Figure 6: Oracle Score Parameters