Technical Reports in Language Technology
Report No 11, 20012
http://www.njas.helsinki.fi/salama

Compounding in English to Swahili machine translation

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Compounding is a common feature in languages. However, the compounding methods differ between languages. These differences must be taken into account when constructing machine translation between two languages. This paper discusses differences of compounding in English and Swahili. It also proposes solutions and demonstrates problem solution using examples for each type of cases.

Key Words: *machine translation, noun compounds, multiword expressions, morphological analysis, disambiguation*

1 Introduction

In English compounding is a common phenomenon. Even the title of this article contains a compound, *machine translation*. A common case is that two nouns form a compound, where the latter member is the general one and the former one qualifies it. Also an adjective and noun can form a compound. Consider the compound *State Minister*, where the general concept *Minister* is specified with the noun *State*. In some cases English allows an alternative structure, where compounding is achieved by using a genitive connector *of*, e.g. *Minister of State*. This applies, however only to certain cases. For example, *traffic police* cannot be converted to **police of traffic*.

More than two nouns can be compounded. Let us see the three-member compound *witch hunting exercise*. Here the hierarchical order in the top-down order is *exercise* > *hunting* > *witch*. It would not be grammatically wrong to express the compound also in forms *exercise of witch hunting* or *exercise of hunting of witches*. However, at least the latter one does not sound fluent English. On the other hand, *exercise of hunting witches* sounds possible. However, the construct is ambiguous, because it can be interpreted in two ways. Either it tells that witch hunting is the content of the exercise, or that the witches that hunt are the target of the exercise. These examples show that noun compounding in English is quite flexible, except in cases where the compound is 'frozen' as a result of frequent use.

How compounds of source language (SL) should be treated in machine translation (MT) depends on the grammar of the target language (TL). Swahili does not allow the

type of compounding, where nouns are just put after each other. Swahili uses the structure, where the genitive connector combines the nouns. This rule applies to structures with two nouns or more. For example, the compound *witch hunting exercise* is translated as *zoezi la uwindaji wa* wachawi (exercise of hunting of witches). However, this general conversion rule does not apply always. There are English compounds that are translated with one word in Swahili.

The translation task contains a number of problems. For example, how do we know which two or more consequent nouns form a compound? And how do we know which noun class prefix should be added to the genitive particle a. It should be selected according to the head noun. But what is the head noun? Is it the preceding noun as in default cases, or is it a more distant noun in the compound as it is in other cases? A further problem is that while in normal compounding, such as 'witch hunting exercise', the noun 'witch' is in singular, although semantically it refers to plural. In Swahili with genitive structure this must be converted to plural.

2 Strategies to handle compounds

There are three strategies to handle noun compounds in English-to-Swahili machine translation, depending on the translation task. The most simple case is the one, where the structure in English and Swahili is the same, such as *Ministry of Health* (Wizara ya Afya). The second case is that the English compound such as *witch hunting* is translated with a genitive structure (uwindaji wa wachawi). In the third type of compounds the translation cannot be constructed on the basis of individual words of the compound. Below we shall handle all these three cases and demonstrate how to proceed in each case for getting the correct translation.

The programming languages and tools used include en-fdg of Connexor, CG-2 by Connexor, Beta, and Perl, and various utilities of Linux.

2.1 Compounds with genitive structure

The translation of compounds with genitive structure is a fairly straight-forward procedure, because the SL and TL allow the same structure.

The text to be translated is: the registration of healers and sellers. The text is analyzed with en-fdg and modified (1).

(1)
```
"<the>"
      "the" %DN> DET
"<registration>"
      "registration" %<P N NOM SG
"<of>"
      "of" %<NOM-OF PREP
"<healers>"
      "healer" %<P N NOM PL
"<and>"
      "and" %CC CC
"<sellers>"
```

```
      "seller" %<P N NOM PL
```

Swahili glosses are added to each word (2).

(2)
```
"<the>"
      "the" %DN> DET
"<registration>"
      "registration" { 11SG sajili } %<P N NOM SG
      "registration" { 11SG sajilishaji } %<P N NOM SG
      "registration" { 11SG sajiri } %<P N NOM SG
"<of>"
      "of" { a } %<NOM-OF PREP
"<healers>"
      "healer" { 1SG 2PL tabibu } %<P N NOM PL
      "healer" { 1SG 2PL uguzi } %<P N NOM PL
      "healer" { 5SG 6PL tabibu } %<P N NOM PL
"<and>"
      "and" { na } %CC CC
"<sellers>"
      "seller" { 1SG 2PL uzaji } %<P N NOM PL
      "seller" { 1SG 2PL uza } %<P N NOM PL
      "seller" { 1SG 2PL sumbaji } %<P N NOM PL
```

The result is disambiguated (3).

(3)
```
"<the>"
      "the" %DN> DET
"<registration>"
      "registration" { 11SG sajili } %<P N NOM SG
"<of>"
      "of" { a } %<NOM-OF PREP
"<healers>"
      "healer" { 1SG 2PL tabibu } %<P N NOM PL
"<and>"
      "and" { na } %CC CC
"<sellers>"
      "seller" { 1SG 2PL uzaji } %<P N NOM PL
```

The correct number in nouns is selected (4).

(4)
```
"<the>"
      "the" %DN> DET
"<registration>"
      "registration" { 11SG sajili } %<P N NOM SG
"<of>"
      "of" { a } %<NOM-OF PREP G-11
"<healers>"
```

```
      "healer" { 2PL tabibu } %<P N NOM PL
"<and>"
      "and" { na } %CC CC
"<sellers>"
      "seller" { 2PL uzaji } %<P N NOM PL
```

The surface form in TL is formed (5).

(5)
```
"<the>"
      "the" %DN> DET
"<registration>"
      "registration" { u+sajili } %<P N NOM SG
"<of>"
      "of" { w+a } %<NOM-OF PREP
"<healers>"
      "healer" { wa+tabibu } %<P N NOM PL
"<and>"
      "and" { na } %CC CC
"<sellers>"
      "seller" { wa+uzaji } %<P N NOM PL
```

No further modification is needed. The word order is as in source language.

2.2 Compounds with consecutive nouns

In English it is possible to form noun compounds simply by sequencing two or more nouns. Swahili allows compounding only with a genitive structure. Let us first examine a compound of two nouns.

The result of analysis is in (6).

(6)
```
"<witch>"
      "witch" %<P N NOM SG
"<hunting>"
      "hunting" %PCOMPL-O N NOM SG
```

Swahili glosses are added (7).

(7)
```
"<witch>"
      "witch" { 1SG 2PL chawi } %<P N NOM SG
      "witch" { 5SG 6PL kahini } %<P N NOM SG
      "witch" { 1SG 2PL numanuma } %<P N NOM SG
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
      "hunting" { 11SG sasi } %PCOMPL-O N NOM SG
```

The text is disambiguated (8).

(8)
```
"<witch>"
      "witch" { 1SG 2PL chawi } %<P N NOM SG
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
```

The structure of the collocation is changed to meet the structure in Swahili (9).

(9)
```
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
"<of>"
      "of" { a } PREP
"<witch>"
      "witch" { 1SG 2PL chawi } %<P N NOM SG COLLOC
```

Select between singular and plural. Note that although *witch* is singular, in Swahili it must be plural (10).

(10)
```
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
"<of>"
      "of" { a } PREP
"<witch>"
      "witch" { 2PL chawi } %<P N NOM SG COLLOC
```

The prefixes are converted to surface form (11).

(11)
```
"<hunting>"
      "hunting" { u+windaji } %PCOMPL-O N NOM SG
"<of>"
      "of" { w+a } PREP
"<witch>"
      "witch" { wa+chawi } %<P N NOM SG COLLOC
```

Now we take a collocation with three nouns. After analysis it is as in (12).

(12)
```
"<a>"
      "a" %DN> DET SG
"<witch>"
      "witch" %<P N NOM SG
"<hunting>"
      "hunting" %PCOMPL-O N NOM SG
```

```
"<exercise>"
      "exercise" %PCOMPL-O N NOM SG
```

Swahili glosses are added (13).

(13)
```
"<a>"
      "a" %DN> DET SG
"<witch>"
      "witch" { 1SG 2PL chawi } %<P N NOM SG
      "witch" { 5SG 6PL kahini } %<P N NOM SG
      "witch" { 1SG 2PL numanuma } %<P N NOM SG
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
      "hunting" { 11SG sasi } %PCOMPL-O N NOM SG
"<exercise>"
      "exercise" { 5SG 6PL zoezi } %PCOMPL-O N NOM SG
      "exercise" { 9SG 10PL tamrini } %PCOMPL-O N NOM SG
```

This is disambiguated (14).

(14)
```
"<a>"
      "a" %DN> DET SG
"<witch>"
      "witch" { 1SG 2PL chawi } %<P N NOM SG
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
"<exercise>"
      "exercise" { 5SG 6PL zoezi } %PCOMPL-O N NOM SG
```

The compound structure is converted to meet the requirements of Swahili (15).

(15)
```
"<a>"
      "a" %DN> DET SG
"<exercise>"
      "exercise" { 5SG 6PL zoezi } %PCOMPL-O N NOM SG
"<of>"
      "of" { a } PREP
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
"<of>"
      "of" { a } PREP
"<witch>"
      "witch" { 1SG 2PL chawi } %<P N NOM SG COLLOC
```

The correct number is selected (16).

(16)
```
"<a>"
      "a" %DN> DET SG
"<exercise>"
      "exercise" { 5SG zoezi } %PCOMPL-O N NOM SG
"<of>"
      "of" { a } PREP
"<hunting>"
      "hunting" { 11SG windaji } %PCOMPL-O N NOM SG
"<of>"
      "of" { a } PREP
"<witch>"
      "witch" { 2PL chawi } %<P N NOM SG COLLOC
```

Prefixes are converted to surface form (17).

(17)
```
"<exercise>"
      "exercise" {  zoezi } %PCOMPL-O N NOM SG
"<of>"
      "of" { l+a } PREP
"<hunting>"
      "hunting" { u+windaji } %PCOMPL-O N NOM SG
"<of>"
      "of" { w+a } PREP
"<witch>"
      "witch" { wa+chawi } %<P N NOM SG COLLOC
```

It should be noted that the conversion from English to Swahili here takes place using general rules. That is, if two English nouns occur after each other, the structure is converted into genitive structure. The genitive connector takes its form from the preceding noun. If three nouns occur after each other, the structure is converted into genitive structure. Each genitive connector takes its form according to the preceding noun.

The above two types of procedures are applied in case no specific rule applies to the case. The third case, where case-specific rules are used, is described below.

3 Various collocations treated as MWEs

There are large numbers of cases, where noun compounds cannot be treated with general rules. The default conversion may fail to produce the correct structure. Even more common is that the gloss defined as a default gloss for each member of the compound does not produce appropriate translation. One solution would be to write case-specific disambiguation rules for problematic words.

However, it is more secure to isolate these cases as MWEs and treat each of them as a single semantic unit. By so doing we make sure that we get the correct translation. By writing case-specific MWE rules it is possible to handle noun compounds and combinations of an adjective and noun, as well as many-to-one MWEs. Rules can be

written with various degrees of strictness. Some rules would not allow any inflection. Some others would allow variation between singular and plural. And in case of verbs, many kinds of inflection would be made possible. Below are some many-to-one types of MWEs.

3 Noun compound with two members

(18)
```
"<identity>"
      "identity" %A> N NOM SG
"<cards>"
      "card" %OBJ N NOM PL
```

This is isolated as a MWE (19).

(19)
```
"<identity>"
      "identity" MW-N>
"<cards>"
      "card" { 8PL tambulisho } N <MW %OBJ N NOM PL
```

Noun compound with genitive structure (20).

(20)
```
"<Member>"
      "member" %PCOMPL-S N NOM SG
"<of>"
      "of" %<NOM-OF PREP
"<Parliament>"
      "parliament" %<P N NOM SG
```

After isolation (21).

(21)
```
"<Member>"
      "member" N NOM SG MW-N>>
"<of>"
      "of" <>MW
"<Parliament>"
      "parliament" N <<MW { 1SG anabunge }
```

A MWE of adjective and noun (22).

(22)
```
"<ruling>"
      "ruling" %A> A ABS
"<party>"
      "party" %A> N NOM
```

After isolation (23).

(23)

```
"<ruling>"
      "ruling" MW-N>
"<party>"
      "party" { 7SG ama-tawala } N <MW %A> N NOM
```

4 Conclusion

There are many kinds of problems in translating English noun compounds into Swahili. We can identify two types of default cases. If English uses genitive structure in compounding, the structure matches with that of Swahili, and no conversion rules are needed. In case the English compound is the type of putting two or more nouns simply after each other, conversion rules must be written. These rules can be, however, general, in the sense that whichever two nouns appearing next to each other will be converted to genitive structure.

Because such general rules are not safe, a large number of case-specific rules are needed. These rules define for each noun compound and other types of MWEs a unique translation in target language. The set of case-specific rules are applied first. Then follow the general rules that convert the English noun compounds - two or more nouns after each other - into structures with genitive connector. The English structures that already are based on genitive connector do not need specific or general rules.

In addition to this translation procedure, there are cases that are ambiguous. These cases should be subjected to disambiguation.