Technical Reports in Language Technology
Report No 12, 2013
http://www.njas.helsinki.fi/salama

Handling proper names in Machine Translation

Arvi Hurskainen

Introduction

Proper names constitute a difficult problem in machine translation (MT). In languages such as Swahili and English, the default is that if a word is written with a capital initial letter, it is a proper name. This holds especially when the word does not begin a sentence. In sentence-initial position any word is written with a capital initial letter, which still complicates the matter.

There is a group of proper names, which need no translation. They can be transferred to the target language (TL) as such. Person names are an example of such names. There are also proper names that have to be translated. Examples of such names are organisation names, that are written with capital initial letters (e.g. *Umoja wa Afrika*), months, and week days. A special group of proper names are the names in the Bible, where also person names must be translated. There is often ambiguity in proper names, because one or more members of a proper name cluster are also ordinary words. For example, *Umoja* means unity in ordinary sense. Texts also abound with person names that should be transferred to the target language as such. However, many of these are also in the list of ordinary words, which is a source of ambiguity.

Approaches to resolve ambiguity

A solution which comes first into mind is to list all proper names into the analysis system, either to the morphological analyser or to a post-processing module. In practice, however, such a list would be always defective, because new names appear continually. A better approach is: list only such names that need translation or such semantic labelling that cannot be determined on the basis of the word form itself. Also the need for adding semantic information may give reason for listing. For example, such distinctions as animate/non-animate, male/female and singular/plural may be reason for listing proper names. For the rest of non-sentence-initial but capital-initial words can be given the default interpretation of proper name. They do not need listing anywhere. If such a word has also an ordinary meaning, the analysis system produces at least two interpretations, one for a proper noun and one or more for ordinary words (note that the word can be a noun or a form of some other word class).

When considering disambiguation, it would be tempting to select the proper name interpretation for words that are capital-initial but not sentence-initial. However, the situation is not that simple. Texts often have capital-initial words inside the sentence, for stylistic and whatever reasons. They should be kept separate from true proper names.

The solution is to mark with a special tag such words that are also potential proper names. Such tags can be added to the morphological lexicon or produce in a post-editing process. Using this method we can control whether the word is a potential proper name. The words that do not have that tag will be interpreted as normal words although they would be capital-initial and non-sentence-initial. This requires keeping control of the list of words that may have both interpretations.

The problem that remains is how to handle ambiguous capital-initial words that are at the same time sentence-initial. In this position all words are capital-initial. It is hardly possible to solve this problem exhaustively. Probability measures would, however, bring satisfactory results.

Implementation

Below I demonstrate the problems discussed above. In (1) we have an analysis result of *'Wakati Rais Jakaya Kikwete akisubiriwa*'. Each word is numbered afterwards to make reference easier.

(1)
1. "<*wakati>"

```
                "wakati" N 11/10-SG { the } { time }  TIME CAP
                "wakati" N 11/10-SG { the } { period of :time }  TIME CAP
                "wakati" N 11/10-SG { the } { point of :time }  TIME CAP
2. "<*rais>"
                "rais" N 9/6-SG { the } { *president } MALE HUM CAP
                "*rais" N TITLE { *president } AN HUM
3. "<*jakaya>"
                "*jakaya" PROPNAME SG { *jakaya }
4. "<*kikwete>"
                "kweta" V INF-SBJN { to } z [kweta] { crawl along }  SV CAP
                "kweta" V INF-SBJN { to } z [kweta] { have a rough :time }  SV CAP
                "kweta" V SBJN 7/8-SG-SP VFIN { it } z [kweta] { crawl along }  SV CAP
                "kweta" V SBJN 7/8-SG-SP VFIN { it } z [kweta] { have a rough :time }  SV CAP
5. "<akisubiriwa>"
                "subiriwa" V 1/2-SG3-SP VFIN { he } COND-IF { if/when } z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { he } COND:ki z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { she } COND-IF { if/when } z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { she } COND:ki z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { he } PR:a 7/8-SG-OBJ OBJ { it } z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { she } PR:a 7/8-SG-OBJ OBJ { it } z [subiri] { wait } SVO PASS
```

We see that most words have more than one interpretation. Word 1 is a sentence-initial normal word and is therefore capital-initial. Word 2 is a normal word and also a title. Both interpretations are listed in the lexicon. Word 3 is only a proper name. It is listed in the lexicon. Word 4 is interpreted only as a verb form of the verb *kweta*. This is the only interpretation of the word *Kikwete*. Word 5 is a normal verb.

We see above that for the word *Kikwete* the correct interpretation is missing. To correct this and many other similar cases we add a tag to *kweta* to show that it is a candidate as a proper name. This is demonstrated in (2).

(2)

```
"<*wakati>"
                "wakati" N 11/10-SG { the } { time }  TIME CAP
                "wakati" N 11/10-SG { the } { period of :time }  TIME CAP
                "wakati" N 11/10-SG { the } { point of :time }  TIME CAP
"<*rais>"
                "rais" N 9/6-SG { the } { *president } MALE HUM CAP
                "*rais" N TITLE { *president } AN HUM
"<*jakaya>"
                "*jakaya" PROPNAME SG { *jakaya }
"<*kikwete>"
                "kweta" V INF-SBJN { to } z [kweta] { crawl along }  SV CAP PROP-MUST
                "kweta" V INF-SBJN { to } z [kweta] { have a rough :time }  SV CAP PROP-MUST
                "kweta" V SBJN 7/8-SG-SP VFIN { it } z [kweta] { crawl along }  SV CAP PROP-MUST
                "kweta" V SBJN 7/8-SG-SP VFIN { it } z [kweta] { have a rough :time }  SV CAP PROP-MUST
"<akisubiriwa>"
                "subiriwa" V 1/2-SG3-SP VFIN { he } COND-IF { if/when } z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { he } COND:ki z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { she } COND-IF { if/when } z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { she } COND:ki z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { he } PR:a 7/8-SG-OBJ OBJ { it } z [subiri] { wait } SVO PASS
                "subiriwa" V 1/2-SG3-SP VFIN { she } PR:a 7/8-SG-OBJ OBJ { it } z [subiri] { wait } SVO PASS
```

The tag PROP-MUST is added to the analysis of Kikwete. Then a rule for adding a proper name interpretation for Kikwete is added and this interpretation is selected if the word is not sentence-initial. Also the general disambiguation is performed.

(3)

```
"<*wakati>"
        "wakati" N 11/10-SG { the } { time }  TIME CAP
"<*rais>"
        "*rais" N TITLE { *president } AN HUM
"<*jakaya>"
        "*jakaya" PROPNAME SG { *jakaya }
"<*kikwete>"
        "*kikwete" PROPNAME { *kikwete }
"<akisubiriwa>"
        "subiriwa" V 1/2-SG3-SP VFIN { he } COND-IF { if/when } z [subiri] { wait } SVO PASS
```

In case *Kikwete* would be the first word of the sentence, this procedure would not bring the correct interpretation. We can resolve the problem on the basis of probability. We can ask: which interpretation is more likely, the subjunctive form of the verb *kweta* (meaning *he may crawl along*) or proper name. The answer in this case is obvious.

More examples

Let us take more examples of well known person names to see how they are analysed by the parsing system. The list includes some names of presidents.

(4)

```
1. "<*julius>"
        "*julius" PROPNAME SG { *julius } MALE
2. "<*nyerere>"
        "nyerere" N 9/10-SG { the } { :copper bangle , :brass bangle } CAP
        "nyerere" N 9/10-PL { the } { :copper bangle , :brass bangle } CAP
3. "<*yakaya>"
        "*yakaya" PROPNAME SG { *yakaya }
4. "<*kikwete>"
        "kweta" V INF-SBJN { to } z [kweta] { crawl along , have a rough :time } SV CAP
        "kweta" V SBJN 7/8-SG-SP VFIN { it } z [kweta] { crawl along , have a rough :time } SV CAP
        "kweta" V INF-SBJN { to } z [kweta] { crawl along , have a rough :time } SV CAP
5. "<*uhuru>"
        "uhuru" N 11-SG { the } { freedom , independence , liberty } CAP
6. "<*kenyata>"
        "*kenyata" PROPNAME SG { *kenyata } MALE
7. "<*mwai>"
        "*mwai" PROPNAME SG { *mwai }
8. "<*kibaki>"
        "baki" V INF-SBJN { to } z [baki] { remain , stay behind , be left behind } SV CAP
        "baki" V SBJN 7/8-SG-SP VFIN { it } z [baki] { remain , stay behind , be left behind } SV CAP
        "baki" V INF-SBJN { to } z [baki] { remain , stay behind , be left behind } SV CAP
        "kibaki" N 7/8-SG { the } { rest , remains } CAP
        "baki" ADV ADV:ki 5/6-SG { the } { remainder , residue , balance } CAP
        "baki" ADV ADV:ki 9/10-SG { the } { remainder , residue and balance , left-overs } CAP
        "baki" ADV ADV:ki 9/10-PL { the } { remainder , residue and balance , left-overs } CAP
9. "<*yoveri>"
        "*yoveri" PROPNAME SG { *yoveri } MALE
10. "<*museweni>"
        "*museweni" PROPNAME SG { *museweni }
11. "<*benjamin>"
        "*benjamin" PROPNAME SG { *benjamin } MALE
12. "<*mkapa>"
        "pa" V 1/2-PL2-SP VFIN { you } NARR:ka z [pa] { give } V SVOO MONOSLB CAP
        "pa" V 18-SG-SP VFIN { there } NARR:ka z [pa] { give } V SVOO MONOSLB CAP
13. "<*ali>"
        "*ali" PROPNAME SG { *ali } MALE
```

```
14. "<*hassan>"
        "*hassan" PROPNAME SG { *hassan } MALE
15. "<*mwinyi>"
        "mwinyi" N 9/6-SG { the } { lord , feudal lord , landlord } MALE HUM CAP
16. "<*karume>"
        "*karume" PROPNAME SG { *karume } MALE
```

Among the 16 names, there are ten which are analysed as proper names. It means that they have no other interpretation in Swahili language. These ten names form two groups. In one group are those (Yakaya, Mwai Museweni) that were not recognised by the analysis system, and because they were capital-initial words, they were interpreted as proper names, which need no translation. In the second group are names (Julius, Kenyata, Yoveri, Benjamin, Ali, Hassan, Karume) that have also the label MALE. This label was inserted in the post-analysis process. The rest of the names (Nyerere, Kikwete, Uhuru, Kibaki, Mkapa, Mwinyi) have also another interpretation. In (5) we see how these words are given the label PROP-CAND to show that they are also candidates for a proper noun.

(5)

```
1. "<*julius>"
        "*julius" PROPNAME SG { *julius } MALE
2. "<*nyerere>"
        "nyerere" N 9/10-SG { the } { :copper bangle }  CAP PROP-CAND
3. "<*yakaya>"
        "*yakaya" PROPNAME SG { *yakaya }
4. "<*kikwete>"
        "kweta" V INF-SBJN { to } z [kweta] { crawl along }  SV CAP PROP-CAND
5. "<*uhuru>"
        "uhuru" N 11-SG { the } { freedom }  CAP
6. "<*kenyata>"
        "*kenyata" PROPNAME SG { *kenyata } MALE
7. "<*mwai>"
        "*mwai" PROPNAME SG { *mwai }
8. "<*kibaki>"
        "baki" V INF-SBJN { to } z [baki] { remain }  SV CAP PROP-CAND
9. "<*yoveri>"
        "*yoveri" PROPNAME SG { *yoveri } MALE
10."<*museweni>"
        "*museweni" PROPNAME SG { *museweni }
11. "<*benjamin>"
        "*benjamin" PROPNAME SG { *benjamin } MALE
12. "<*mkapa>"
        "pa" V 1/2-PL2-SP VFIN { you } NARR:ka z [pa] { give } V SVOO MONOSLB CAP PROP-CAND
13. "<*ali>"
        "*ali" PROPNAME SG { *ali } MALE
14. "<*hassan>"
        "*hassan" PROPNAME SG { *hassan } MALE
15. "<*mwinyi>"
        "mwinyi" N 9/6-SG { the } { lord }  MALE HUM CAP PROP-CAND
16. "<*karume>"
        "*karume" PROPNAME SG { *karume } MALE
```

Now when the ambiguous words are labelled as candidates for proper names (PROP-CAND), a rule can be written with the constraint that if a word with the label PROP-CAND is not sentence-initial, interpret it as a proper name. The result is in (6).

(6)

```
 1. ( "<*julius>" "*julius" PROPNAME SG { *julius } MALE  )
 2. ( "<*nyerere>" PROPNAME { *nyerere } )
 3. ( "<*yakaya>" "*yakaya" PROPNAME SG { *yakaya } )
```

4. ("<*kikwete>" PROPNAME { *kikwete })
5. ("<*uhuru>" PROPNAME { *uhuru })
6. ("<*kenyata>" "*kenyata" PROPNAME SG { *kenyata } MALE)
7. ("<*mwai>" "*mwai" PROPNAME SG { *mwai })
8. ("<*kibaki>" PROPNAME { *kibaki })
9. ("<*yoveri>" "*yoveri" PROPNAME SG { *yoveri } MALE)
10. ("<*museweni>" "*museweni" PROPNAME SG { *museweni })
11. ("<*benjamin>" "*benjamin" PROPNAME SG { *benjamin } MALE)
12. ("<*mkapa>" PROPNAME { *mkapa })
13. ("<*ali>" "*ali" PROPNAME SG { *ali } MALE)
14. ("<*hassan>" "*hassan" PROPNAME SG { *hassan } MALE)
15. ("<*mwinyi>" PROPNAME { *mwinyi })
16. ("<*karume>" "*karume" PROPNAME SG { *karume } MALE)

All the names in (6) have a proper name interpretation, also the first one. The reason is that Julius has anly a proper name interpretation. In (7) we see what happens when we have an ambiguous name as the first name.

(7)
1. ("<*nyerere>" "nyerere" N 9/10-SG { the } { :copper bangle } CAP)
2. ("<*julius>" "*julius" PROPNAME SG { *julius } MALE)
3. ("<*yakaya>" "*yakaya" PROPNAME SG { *yakaya })
4. ("<*kikwete>" PROPNAME { *kikwete })
5. ("<*uhuru>" PROPNAME { *uhuru })
6. ("<*kenyata>" "*kenyata" PROPNAME SG { *kenyata } MALE)
7. ("<*mwai>" "*mwai" PROPNAME SG { *mwai })
8. ("<*kibaki>" PROPNAME { *kibaki })
9. ("<*yoveri>" "*yoveri" PROPNAME SG { *yoveri } MALE })
10.("<*museweni>" "*museweni" PROPNAME SG { *museweni })
11.("<*benjamin>" "*benjamin" PROPNAME SG { *benjamin } MALE)
12.("<*mkapa>" PROPNAME { *mkapa })
13. ("<*ali>" "*ali" PROPNAME SG { *ali } MALE)
14. ("<*hassan>" "*hassan" PROPNAME SG { *hassan } MALE)
15. ("<*mwinyi>" PROPNAME { *mwinyi })
16. ("<*karume>" "*karume" PROPNAME SG { *karume } MALE)

The reason that Nyerere was not interpreted as a proper name is that it is too risky to interpret the sentence-initial ambiguous word as a proper name, because all sentence-initial words are capital-initial, and we easily interpret wrong words as proper names. Sentence-initial proper name candidates need more fine grained treatment with specific rules.

Conclusion

We may sum up the above discussion on handling proper names.
1. List often occurring proper names in the morphological lexicon or describe them in a post processing phase.
2. Pruvide with a special tag such words that may appear as ordinary words or as proper names.
3. Write a rule set for the following algorithm: If the word has the special tag and the word is capital-initial and non-sentence-initial, interpret it as a proper name and remove the original meaning.
4. In case the word already has a proper name interpretation in addition the ordinary word interpretation, disambiguate it using the normal disambiguation routines.
5. If the ambiguous word is sentence-initial, treat it with fine-grained detailed rules.