Optimizing disambiguation in Swahili

Arvi HURSKAINEN
Institute for Asian and African Studies
Box 59
FI-00014 University of Helsinki
Arvi.Hurskainen@helsinki.fi

Abstract

It is argued in this paper that an optimal solution to disambiguation is a combination of linguistically motivated rules and resolution based on probability or heuristic rules. By disambiguation is here meant ambiguity resolution on all levels of language analysis, including morphology and semantics. The discussion is based on Swahili, for which a comprehensive analysis system has been developed by using two-level description in morphology and constraint grammar formalism in disambiguation. Particular attention is paid to optimising the use of different solutions for achieving maximal precision with minimal rule writing.

1 Introduction

In ambiguity resolution of natural language, both explicit linguistic information and probability calculation have been used as basic approaches. In early experiments usually only one strategy was applied, so that ambiguity resolution was performed either with the help of linguistic rules, or through probability calculation. Advanced approaches make use of both strategies, and they differ mainly in what kind of role each of these two methods has in the system (Wilks and Stevenson 1998, Stevenson and Wilks 2001). Sources of structured data, such as the WordNet (Miller 1990; Resnik 1998b; Banerjee and Pedersen 2002), have also been made use of.

It is commonly known that the more comprehensive the description of language is, the more ambiguous is the interpretation of individual words[1]. Ambiguity occurs between word classes, between variously inflected word-forms, and above all, between various meanings of a word. A fairly large number[2] of words in different word categories have more than one clearly distinguished meaning. Semantic disambiguation tends to be the hardest part of the disambiguation process, largely because of the fact that in semantics there are few distinguishable categories that could be used for generalising disambiguation rules. Below I shall describe a method where the use of linguistic rules and probability has been optimised with minimal loss of linguistic precision.

Morphological description is carried out in the framework of two-level formalism[3]. After having been under development for 19 years (Hurskainen 1992, 1996), the parser of Swahili has now reached a phase where the recall as well as the precision[4] is close to 100% in unrestricted standard Swahili text.

---

[1] By word is here meant any string of characters, excluding punctuation marks and diacritics. Also, multi-word concepts, if they are handled as single entities, are considered words.

[2] I do not consider it meaningful to present statistical details of ambiguity, because, when semantic glosses are included, the borderline between real ambiguity and such ambiguity as is found between synonyms and near-synonyms is vague.

[3] The development environment for designing the morphological parser was provided by Lingsoft and Kimmo Koskenniemi (1983).

[4] The criterion of precision in morphological analysis is considered fulfilled if one of the readings of a word is correct in the context concerned, and all other readings are grammatically correct analyses in some other context.

Disambiguation rules, as well as the rules for syntactic mapping (not discussed here) and for identifying idioms, were written within the framework of constraint grammar by using the CG-2 parser[5]. In other words, morphological disambiguation and semantic disambiguation were implemented within a single rule system. This was possible because the CG-2 parser treats all strings in the analysis result, including glosses in English, as tags that can be made use of in rule writing (Tapanainen 1996: 6).[6]

The properties of the CG-2 parser include the following:

(a) With a rule one may either select or remove a reading from a cohort[7].

(b) The application of a rule can be constrained in several ways by making use of the occurrence or absence of features. Reference to the position of the constraining feature can be precisely made forwards and backwards within the sentence.

(c) The identification of constraining features can be made relational by more than one phase of scanning, whereby after finding one feature, scanning may be continued again in either direction. By default, scanning terminates at a sentence boundary, but its termination can also be defined elsewhere.

(d) Rule conditions can be expressed either directly with concrete tags or indirectly by using set names. The latter facility simplifies rule writing, especially of general rules.

(e) The possibility of concatenating tag sets as well as concrete tags decreases considerably the need of defining tag sets.

(f) The application of rule order can be defined by placing the rules into sections, so that the more general and reliable rules come first and other rules later in the order of decreasing reliability. This also makes it possible to write heuristic rules within the same rule system.

(g) Mapping rules, which are the standard rules for syntactic mapping, also include a possibility of adding a new reading as well as of replacing the reading of a line. The latter facility is demonstrated below when discussing idioms.

2 Maximal morphological and semantic description as precondition

The basic strategy in processing is that the morphological description is as full and detailed as possible. Each string in text is interpreted and all possible interpretations of each string are made explicit. The maximal recall and precision are achieved by updating the dictionary from time to time with the help of the changing target language[8]. As a result of analysis there is a text where every string has at least one interpretation and no legitimate interpretation is excluded. Example (1) illustrates the point.

(1)

```
Kiboko
   "kiboko"  N 7/8-SG  { fat person } HUM
   "kiboko"  N 7/8-SG  { whip , strip of hippo hide }
   "kiboko"  N 7/8-SG  { hippo , hippopotamus } AN
   "kiboko"  N 7/8-SG
      { beautiful/attractive/outstanding thing }
   "kiboko"  N 7/8-SG  { ornamental stitch }
   "boko"  ADV ADV:ki  9/10-SG  { gourd for
      drinking water or local brew }
   "boko"  ADV ADV:ki  9/10-PL  { gourd for
      drinking water or local brew }
aishiye
   "ishi"  V 1/2-SG3-SP VFIN { live , reside , stay }
      SV AR GEN-REL 1/2-SG
kwenye
   "kwenye" PREP { in , at }
   "enye" PRON 15-SG { which has }
   "enye" PRON 17-SG { place which has }
maziwa
```

[5] The environment for writing and testing disambiguation rules was provided by Connexor and Pasi Tapanainen (1996).

[6] In disambiguation, the precision criterion is considered fulfilled if the reading chosen in that context is correct. In two independent tests with recent news texts of 5,000 words each, the precision was 99.8% and 99.9%.

[7] A cohort is a word-form plus all its morphological interpretations.

[8] By target language I mean the kind of text, for which the application is intended. It is hardly possible to maintain a dictionary that is optimal for handling all types of domain-specific texts. Although the large size of the dictionary would not be a problem, it would be difficult to handle e.g. such words that in one type of text are individual lexemes but in another domain are part of multi-word concepts that should be treated as one unit. In addition to new words, misspellings also cause problems. Some commonly occurring misspellings and non-standard spellings can be encoded into the dictionary and thus give the word a precise interpretation.

```
   "ziwa"  N 5a/6-PL  { lake }
   "ziwa"  N 5a/6-PL  { breast }
   "maziwa" N 6-PL  { milk }
amekula
   "la"  V 1/2-SG3-SP VFIN PERF:me 1/2-SG2-OBJ
      OBJ { eat } SV SVO MONOSLB
   "la"  V 1/2-SG3-SP VFIN PERF:me 15-SG-OBJ
      OBJ { eat } SV SVO MONOSLB
   "la"  V 1/2-SG3-SP VFIN PERF:me 17-SG-OBJ
      OBJ { eat } SV SVO MONOSLB
   "la"  V 1/2-SG3-SP VFIN PERF:me INFMARK
     { eat } SV SVO MONOSLB
nyanya
   "nyanya"  N 5a/6-SG  { tomato }
   "nyanya"  N 9/10-SG  { tomato }
   "nyanya"  N 9/10-SG  { grandmother } HUM
   "nyanya"  N 9/10-PL  { tomato }
   "nyanya"  N 9/10-PL  { grandmother } HUM
   "nyanya"  N 9/6-SG  { grandmother }  HUM
.$
```

Without disambiguation, the following interpretations are possible:
(a) A fat person, who lives in lakes, has eaten tomatoes. (b) A fat person, who lives in lakes, has eaten grandmothers. (c) A fat person, who lives in breasts, has eaten tomatoes. (d) A fat person, who lives in breasts, has eaten grandmothers. (e) A fat person, who lives in milk, has eaten tomatoes. (f) A fat person, who lives in milk, has eaten grandmothers. (g) A hippo, which lives in lakes, has eaten tomatoes. (h) A hippo, which lives in lakes, has eaten grandmothers. (i) A hippo, which lives in breasts, has eaten tomatoes. (j) A hippo, which lives in breasts, has eaten grandmothers. (k) A hippo, which lives in milk, has eaten tomatoes. (l) A hippo, which lives in milk, has eaten grandmothers.

The situation would be even worse if "aishiye" with relative marker (GEN-REL 1/2-SG) were missing. It requires that the preceding referent be animate and thus excludes inanimate alternatives. The subject prefix in the main verb "amekula" also refers to an animate subject. But because it can also stand without an overt subject, this clue is not reliable.

When we look for the possible subject in the sentence, we seem to have three candidates. "Kiboko" certainly is one of them, because it is a noun and some of its readings agree[9] with the subject prefix of the main verb. In regard to its position, "ziwani" would also suit, but it is ruled out because it has a locative suffix. Finally, no overt subject would be necessary, whereby the phrase preceding the main verb would be an object dislocated to the left and the sentence would mean, "The grandmother has eaten the hippo/fat person who lives in the lakes/breasts/milk".

3 Disambiguation with linguistic rules

From the analysed sentence we can see that part of the ambiguity is easy to resolve with rules. For example, "kiboko" cannot be an adverbial form (ADV:ki) of "boko" (= in the manner of a gourd), because it is the referent of the following relative verb "aishiye", which for its part requires that the referent has to be animate. Therefore, the interpretation "whip" and more rare meanings, "beautiful thing" and "ornamental stitch", are also ruled out. So we are left with two animate meanings, "fat person" and "hippo", for which there are no reliable tags available for writing disambiguation rules.

One of the three interpretations of "kwenye" can be removed (15-SG), because no infinitive precedes it. The word "maziwa" with three interpretations has no grammatical criteria for disambiguation.

The interpretations with object marker (OBJ) of "amekula" (has eaten you) can be removed on the basis of the following noun (without locative), which is reliably the real object.

For "nyanya" there are no reliable criteria for disambiguation. Because it is in object position and without qualifications, no clues for disambiguation can be found among agreement markers.

After applying linguistic disambiguation rules[10], we have an analysis as in (2).

(2)
```
Kiboko
   "kiboko"  N 7/8-SG  { fat person } HUM
   "kiboko"  N 7/8-SG  { hippo , hippopotamus }
```

[9] In this case agreement means something other than morphological agreement. The noun belongs to Class 7 (7/8-SG) and the subject prefix of the verb to Class 1 (1/2-SG3-SP), but the semantic principle, i.e. animacy, overrides the formal criterion.

[10] Because of space restrictions, those rules are not reproduced here.

```
      AN AR
aishiye
   "ishi"  V 1/2-SG3-SP VFIN { live ,  reside , stay }
      SV GEN-REL 1/2-SG
kwenye
   "kwenye" PREP { in , at }
   "enye" PRON 17-SG { place which has }
maziwa
   "ziwa"  N 5a/6-PL  { lake }
   "ziwa"  N 5a/6-PL  { breast }
   "maziwa" N 6-PL  { milk }
amekula
   "la"  V 1/2-SG3-SP VFIN PERF:me INFMARK
      { eat } SV SVO MONOSLB
nyanya
   "nyanya"  N 5a/6-SG  { tomato }
   "nyanya"  N 9/10-SG  { tomato }
   "nyanya"  N 9/10-SG  { grandmother } HUM
   "nyanya"  N 9/10-PL  { tomato }
   "nyanya"  N 9/10-PL  { grandmother } HUM
   "nyanya"  N 9/6-SG  { grandmother }  HUM
.$
```

4 Disambiguation with context-sensitive semantic rules

Now follows the hard part of disambiguation, because no reliable linguistic rules can be written. The easiest case is "kwenye", because the two interpretations represent different phases of the grammaticalization process, and the semantic difference between them is marginal. The preposition "kwenye" is in fact formally a locative (17-SG) form of the relative word "enye" (which has).

For "Kiboko" we can make use of the common knowledge that fat persons do not normally live in lakes, or in breasts, or in milk. Therefore, a rule based on the co-occurrence of "kiboko" and "maziwa"[11] with appropriate meanings can be written.

The word "maziwa" is even more difficult to disambiguate. The word "kiboko" in the sense of hippo can easily co-occur with all three meanings of "maziwa". Here we have to rely on probability[12].

[11] A set of words referring to places where a hippo resides can be defined and used in the rule.

[12] It is possible to write also a context-sensitive rule, where use is made of the fact that rhinos can live in lakes but not in breasts or milk, but such a rule easily becomes too specific.

The word "nyanya" in object position is almost impossible to disambiguate elegantly. The subject of eating can be one or more tomatoes, as well as one or more grandmothers. It is not rare at all that hippos devour people, although there is no proof that they would be particularly fond of grandmothers. Nobody has heard fat men eating grandmothers, but those do not come into question in any case, because they do not live in lakes.

If we assume that hippos hardly eat grandmothers we can remove the reading, which has the tag "grandmother". We are still left with singular and plural alternatives of tomato. Here plural would be more natural, because tomatoes are here treated as a mass rather than as individual fruits.

When context-sensitive semantic rules and heuristic rules are applied, the reading is as shown in (3).

(3)

```
Kiboko
   "kiboko"  N 7/8-SG  { hippo , hippopotamus } AN
aishiye
   "ishi"  V 1/2-SG3-SP VFIN { live , reside , stay }
      SV GEN-REL 1/2-SG
kwenye
   "kwenye" PREP { in , at }
maziwa
   "ziwa"  N 5a/6-PL  { lake }
amekula
   "la"  V 1/2-SG3-SP VFIN PERF:me INFMARK
      { eat } SV SVO MONOSLB
nyanya
   "nyanya"  N 9/10-PL  { tomato }
.$
```

5 Problem of semantic generalisation

Although the possibilities for generalisation in semantics are limited, in noun class languages relevant semantic clusters can be found. Even though classes in Swahili are only in exceptional cases semantically 'pure', the class membership often provides sufficient information for disambiguation, either by direct selection or, more often, by exclusion of a reading.

The grades of animacy (e.g. human, animal, vegetation) are an example of useful semantic groupings, which can be used in generalising disambiguation. Another useful feature, actually belonging to syntax, is the division of verbs into

categories according to their argument structure (e.g. SV, SVO, SVOO)

Neural networks have been used successfully for identifying clusters of co-occurrence of words and their accompanying tags (Veronis and Ide 1990; Sussna 1993; Resnik 1998a). Research results, carried out with the Self-Organizing Map (Kohonen 1995) on semantic clustering of verbs and their arguments in Swahili, are very promising, and useful generalizations have been found (Ng'ang'a 2003).[13] These findings can be encoded into the morphological parser and used in writing semantic disambiguation rules.

6 When means for rule writing fail

It sometimes happens that linguistic disambiguation rules cannot be written. Particularly problematic is the noun of the Class 9/10 in object position without qualifiers, many of which would help in disambiguation. In this noun class there are no features in nouns for determining whether the word is in singular or plural[14]. The detailed survey of about 11,000 occurrences of class 9/10 nouns in object position shows, however, that 97% of them are unambiguously in singular. Among the remaining 3%, 2% can be either in singular or plural, and only one percent are such cases where the noun is clearly in plural. These 2% are typically count nouns, which sometimes can be disambiguated, if, for example, they are members in a list of nouns. Nouns in such lists tend to be either in singular or in plural, and often at least one list member belongs to one of the other noun classes, where singular and plural are distinguished.

The solution for the nouns of the class 9/10 in object position is, therefore, that for the rare plural cases, disambiguation rules are written, while singular is the default interpretation.

[13] The likelihood of co-occurrence can be established between word pairs, or clusters, and also between words and tags attached to them. Therefore, the full range of information in an analysed corpus can be utilized in establishing relationships.

[14] Singular and plural are identical in this class, and it is the biggest class of the language, consisting of about 39% of all nouns.

7 Treatment of multi-word concepts and idioms

In computational description of a language, multi-word concepts and idioms can be treated as one unit, because in both cases the meaning is based on more than one string in text. If a multi-word concept consists of a collocation or noun phrase, it can be encoded in the tokenizer (4) and the morphological lexicon (5). Such constructions have two forms (SG and PL) at the most.

(4) bwana shamba > bwana_shamba
jumba la makumbusho > jumba_la_makumbusho
majumba ya makumbusho >
majumba_ya_makumbusho

(5) bwana_shamba
"bwana_shamba" N 9/6-SG { agricultural adviser}
HUM
jumba_la_makumbusho
"jumba_la_makumbusho" N 5/6-SG { museum }
majumba_ya_makumbusho
"majumba_ya_makumbusho" N 5/6-PL
{ museums}

If the concept has a non-finite verb as part of the construction, as is often the case in idioms, the constructions cannot be handled on the surface level. It is possible to handle them with disambiguation rules. Example (6), which is an idiom, shows how each of its constituent parts is interpreted in isolation.

(6)
alipiga
"piga" V 1/2-SG3-SP VFIN PAST { hit , beat }
SVO
konde
"konde" N 5/6-SG { cultivated land , fist}
la
"la" GEN-CON 5/6-SG { of }
nyuma
"nyuma" ADV { behind }

With the help of disambiguation rules, the idiom can be identified, although the verb "piga" may have several surface forms, including extended forms. The solution adopted here is the following:

As a first step we identify the constituent parts of the idiom and describe its structure by a tag, as is shown in (7). The angle brackets (<>>) show that the idiom contains the current word as well as the preceding word and two following words.

Also the meaning of the idiom ("to bribe") is attached to this word.

(7)

```
alipiga
   "piga"  V 1/2-SG3-SP VFIN PAST { hit , beat }
      SVO
konde
   "konde"  <>>IDIOM { to bribe }
la
   "la"  GEN-CON 5/6-SG { of }
nyuma
   "nyuma"  ADV { behind }
```

Then we mark each of the other constituent parts of the idiom and show their relative location in the structure by using angle brackets, as shown in (8). For example, "nyuma" is the last constituent and all three words before it are part of the idiom. Original glosses of other constituent parts are removed. The verb retains its morphological tags, and a special tag (IDIOM-V) is added to show that it is part of the idiom.

(8)

```
alipiga
   "piga"  V 1/2-SG3-SP VFIN PAST SVO  IDIOM-V
konde
   "konde" <>>IDIOM { to bribe }
la
   "la" IDIOM<<>
nyuma
   "nyuma" IDIOM<<<
```

8 Making use of default interpretation

Although it would be possible to write disambiguation rules for practically all such cases where sufficient features for rule writing are available, it is sometimes impractical, especially in selecting the right semantic interpretation. This can be implemented in more than one way, for example by constructing the morphological analyser so that the alternative semantic analyses are in frequency order (9).

(9)

```
taa
   "taa"  N 9/10-SG  { lamp , lantern } AR
   "taa"  N 9/10-SG  { discipline , obedience }
   "taa"  N 9/10-SG  { large flat fish , skate } AN
   "taa"  N 9/10-PL  { lamp , lantern } AR
   "taa"  N 9/10-PL  { discipline , obedience }
   "taa"  N 9/10-PL  { large flat fish , skate } AN
```

The word "taa" gets three semantic interpretations, each in singular and plural. The most obvious gloss (lamp, lantern) is the first in order, and if no rule has chosen any of the other alternatives, this one is chosen as the default case. The choice of other alternatives is controlled by rules as far as possible. For example, the animate reading can often be chosen with congruence rules.

9 Discussion

The disambiguation of a language is a process where the cooperation of linguistic rules and probability should be optimised. It was shown above briefly that different disambiguation operations should be cascaded so that the most reliable disambiguation is carried out first and the least reliable cases last. Multi-word concepts can be handled so that such constructions that do not have inflecting constituent parts are treated as part of morphology, and those with inflecting parts, especially idioms, are handled with disambiguation rules. We have also seen that linguistic rules should precede rules based on probability. It is also possible to simplify the writing of semantic rules by constructing the morphological parser so that semantic readings come in order of frequency, whereby the most frequent interpretation is considered a default case, and only other interpretations need rules. The experiments with the SOM algorithm indicate that it is possible to find significant relationships between adjacent words on the one hand and between words and tags on the other. Such information can then be encoded in the morphological dictionary and used in generalising disambiguation rules. Ambiguity resolution can be enhanced further by constructing explicit dependencies between constituent parts of a sentence (Järvinen and Tapanainen 1997; Tapanainen and Järvinen 1997; Tapanainen 1999) or by making use of a parse tree bank of the type of WordNet (Hirst and Onge 1998).

10 Acknowledgements

Thanks go to Lingsoft and Kimmo Koskenniemi for allowing me to use the Two-Level Compiler for handling morphological analysis and to Connexor and Pasi Tapanainen for prividing access to CG-2 for writing disambiguation rules.

References

Banerjee, S. and T. Pedersen, 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, pp. 136-145.

Fellbaum, C. (Ed.) 1998. *WordNet: An electronic lexical database*. MIT Press.

Hirst, G. and D. St. Onge 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In: C. Fellbaum (Ed.), *WordNet: An electronic lexical database*. MIT Press, pp. 305-332.

Hurskainen A. 1992. A Two-Level Computer Formalism for the Analysis of Bantu Morphology. An Application to Swahili. *Nordic Journal of African Studies* 1(1): 87-122.

Hurskainen A. 1996. Disambiguation of morphological analysis in Bantu languages. In: *Proceedings of COLING-96,* pp. 568-573.

Järvinen, T. and P. Tapanainen, 1997. A *Dependency Parser for English.* Technical Reports, No. TR-1. Department of General Linguistics, University of Helsinki.

Kohonen, T. 1995. *Self-Organizing Maps*. Berlin: Springer.

Koskenniemi, K. 1983. *Two-level morphology: A general computational model for word-form recognition and production.* Publications No.11. Department of General Linguistics, University of Helsinki.

Miller, G. 1990. *Wordnet: An On-line Lexical Database*. International Journal of Lexicography 3(4): 235-312.

Ng'ang'a, J. 2003. *Semantic Analysis of Kiswahili Words Using the Self Organizing Map*. Nordic Journal of African Studies, 12(3): 407-425.

Resnik, P. 1998a. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11:95-130.

Resnik, P. 1998b. WordNet and class-based probabilities. In: Fellbaum (Ed.), *WordNet: An electronic lexical database*. MIT Press, pp. 239-263.

Stevenson, M. and Y. Wilks, 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics* 27(3): 321-349.

Sussna, M. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In: *Proceedings of the Second International Conference on Information and Knowledge Management*, pp. 67-74.

Tapanainen, P. 1999. *Parsing in two frameworks: finite-state and functional dependency grammar*. Ph.D. thesis, Department of General Linguistics, University of Helsinki.

Tapanainen, P. 1996. *The Constraint Grammar Parser CG-2.* Publications No. 27. Department of General Linguistics, University of Helsinki.

Tapanainen, P. and T. Järvinen, 1997. A non-projective dependency parser. *ANLP'97*, Washington, pp. 64-71.

Veronis, J. and N. Ide, 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In: *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, pp. 389-394.

Wilks, Y. and M. Stevenson, 1998. Word sense disambiguation using optimised combinations of knowledge sources. In: *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 1398-1402.