Word Sense Disambiguation of Swahili:
Extending Swahili Language Technology with Machine Learning

Wanjiku NG'ANG'A

Academic Dissertation

To be publicly discussed, by due permission of the Faculty of Arts of the University of Helsinki, in auditorium XII, Unioninkatu 34, on the 18th of November, 2005, at 10 o'clock

University of Helsinki
Department of General Linguistics
P.O. Box 9
FIN-00014 University of Helsinki
Finland

PUBLICATIONS
No. 39
2005

"Empowering African Languages in the Information age"

ISSN 0355-7170
ISBN 952-10-2757-6 (paperback)
ISBN 952-10-2758-4 (PDF)
http://ethesis.helsinki.fi

Abstract

This thesis addresses the problem of word sense disambiguation within the context of Swahili-English machine translation. In this setup, the goal of disambiguation is to choose the correct translation of an ambiguous Swahili noun in context. A corpus based approach to disambiguation is taken, where machine learning techniques are applied to a corpus of Swahili, to acquire disambiguation information automatically. In particular, the Self-Organizing Map algorithm is used to obtain a semantic categorization of Swahili nouns from data. The resulting classes form the basis of a class-based solution, where disambiguation is recast as a classification problem. The thesis exploits these semantic classes to automatically obtain annotated training data, addressing a key problem facing supervised word sense disambiguation. The semantic and linguistic characteristics of these classes are modelled as Bayesian belief networks, using the Bayesian Modelling Toolbox. Disambiguation is achieved via probabilistic inferencing. The thesis develops a disambiguation solution which does not make extensive resource requirements, but rather capitalizes on freely-available lexical and computational resources for English as a source of additional disambiguation information. A semantic tagger for Swahili is created by altering the configuration of the Bayesian classifiers. The disambiguation solution is tested on a subset of unambiguous nouns and a manually created gold standard of sixteen ambiguous nouns, using standard performance evaluation metrics.

Preface and Acknowledgements

Growing up in Kenya, a multicultural country where it is the norm rather than the exception to speak several languages, I was intrigued by the diversity of language and at the same time by its universality, from an early age. It was however, at the Institute of Computer Science, University of Nairobi, that my mind was opened up to the world of language technology, after attending a course on Semantic Networks and Natural Language Understanding. I was truly fascinated by the prospects of building computer systems that could understand natural language with all its intricacies. I am therefore very grateful to my first teachers, Dr. Katherine Getao of the University of Nairobi for exposing me to the field of natural language processing, and to Professor Stephen Pulman formerly of Cambridge University, for inspiring and encouraging me to pursue further work in this field.

I am most grateful to my advisors at the University of Helsinki – Professor Arvi Hurskainen and Professor Lauri Carlson for valuable feedback on my work, for their sure commitment to my progress and for securing funding for my research. I am indebted to Prof Arvi Hurskainen for providing me with the opportunity to carry out my research within the Swahili project. Worthy of special mention is the immense contribution that his $20+$ years of research have made to Swahili Language technology, and without which, the language resources required to complete this work would have been unavailable. I am immensely grateful to Professor Lauri Carlson for taking an interest in my work and for sharing his immense knowledge in multilingual semantic analysis and translation with me. His constant encouragement, coupled with his easily-approachable nature, gave me the confidence I needed to complete this work. In addition, He and his wife Seija and family gave me a rare

and valuable opportunity to enjoy true Finnish hospitality on several occasions. *Kiitos paljon*!

I would also like to sincerely thank my pre-examiners, Professor Sonja Bosch (University of South Africa) and Professor Timo Honkela (Helsinki University of Technology) for their effort and valuable comments on my work.

I wish to register my appreciation to the Complex Systems Computation Group at the Helsinki Institute for Information Technology for giving me an audience in the early stages of my research and for their valuable feedback on my ideas. I am especially indebted to Tomi Silander who so generously shared of his vast knowledge of Bayesian learning and gave of his time to answer my numerous queries and provide very useful feedback. I am very grateful too, for providing me with the Bayesian Modelling toolbox which has been used extensively in this work.

I would like to thank the following people from the department of Linguistics and the Institute for Asian and African studies for their varied support over the course of my research: Professor Kimmo Koskenniemi, Hannah Westerlund, Harry Halén, Juri Ahlfors, Minttu Hurme, Ekaterina Gruzdeva, Orvokki Heinämäki for valuable discussions on general linguistics and friendly encouragement, Krister Lindén for lively discussions on machine learning and for valuable comments on sections of my draft and Professor Fred Karlsson for valuable advice.

To Dr. Ike Achebe, Dr. Robert Gateru, Filbert Musau, Jean de Dieu Karangwa, Ndung'u Mbagi, Nana Yaw Fosu, Peter Maribei, Malikha Shambe, Matu Mugo, Robert Rogoi, Wathira Ngugi, Caroline Muthoni, Dr. Gregory Kamwendo, Sunny Mbabazi, the *Waimbaji* choir and to all my friends, I say *shukrani* for your friendship and support.

I wish to thank all my family members for their love and prayers and especially to my cousin Nduta Kiragu for being 'electronically' there for me at all times, and to my uncle Professor Ngugi for setting a great example for me to follow. My

deepest gratitude goes to my dearest HB, Kent Libiso, for loving me, believing in me and for praying with me. His unwavering support was crucial for the successful completion of this work. To my dearest sisters – Njeri, Wairimu and Njoki – words cannot express how much your love and support has meant to me over these long years! A big thank you to my nieces Wambui, Njeri, Wanjiku and Wambui and to my nephews Githinji and Ng'ang'a for coping understandingly with my absence. I am extremely grateful to my parents for always encouraging me to go after my dreams, for providing me with all the support, love and prayers that have seen me through all my life and most importantly for their deep faith and conviction that has been a source of strength and hope to me. I would not have accomplished this work without their support. I also wish to thank mum's group of women friends for their constant concern, words of encouragement and prayers.

Last and most importantly, to the Almighty, for granting me good health, peace of mind and mental strength to cope and persevere through it all.

Dedication

To my first instructors, my parents,Wambui and Ng'ang'a.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AI Artificial Intelligence

BBN Bayesian Belief Networks

BL Bayesian Learning

BMT Bayesian Modelling Toolbox

CLIR Cross-lingual Information Retrieval

CPT Conditional Probability Table

IE Information Extraction

IR Information Retrieval

LT Language Technology

MBC Model Based Clustering

MFS Most Frequent Sense

MI Mutual Information

ML Machine Learning

MRD Machine Readable Dictionaries

MT Machine Translation

MAP Maximum a Posteriori

NB Naïve Bayes

NLP Natural Language Processing

SOM Self Organizing Map

SWACGP Swahili Constraint Grammar Parser

SWATWOL Swahili Two-Level Parser

TWS Target Word Selection

WSD Word Sense Disambiguation

Chapter 1
Introduction

The information age has been characterized by the development and convergence of computing, telecommunications and multilingual information systems. This has resulted in the availability of enormous volumes of information in electronic media, but whose natural language form, unlike the data presentation formats typical of computer systems in the past, is more suited for human users than computer systems. This has prompted the development of technologies that would solve this problem and support faster and more efficient access to this information. Natural Language Processing (NLP) provides tools and techniques that facilitate the implementation of natural language-based interfaces to computer systems, enabling communication in natural languages between man and machine. These techniques also enable people to organize, extract and use the knowledge contained in these huge collections of natural language electronic data. Examples of Language Technology (LT) applications include Machine Translation (MT), Information Extraction (IE), Information Retrieval (IR), document classification and summarization, speech recognition and synthesis, to name a few.

However, a pervasive problem afflicting most LT applications is that of ambiguity. Many words have more than one meaning, depending on the context of use. The process by which the most appropriate meaning of an occurrence of an ambiguous word is determined is known as Word Sense Disambiguation (WSD), and remains an open problem in NLP. For humans, resolving ambiguity is a routine task that hardly requires conscious effort. In addition to having a deep understanding of language and its use, humans possess a broad and conscious understanding of the real world, and this equips them with the knowledge that is relevant to make sense disambiguation decisions effortlessly, in most cases. However, creating extensive knowledge-bases which can be used by computers to 'understand' the world and reason about word meanings accordingly, is still an unaccomplished goal of Artificial Intelligence (AI). Consequently, approaches to automatic WSD mainly focus on knowledge-lean methods.

With the availability of huge computer-readable text corpora and the corresponding development of statistical techniques for data mining, corpus-based methods have taken centre stage in the development of WSD solutions. These methods have been employed in the learning of probabilistic models for WSD from large collections of natural language texts. Probabilistic models for sense classification consist of feature variables, the class variable and a probability distribution that models the interactions amongst all the variables. The context of an ambiguous word is defined very simply and usually consists of linguistic information that can be easily extracted from the neighbourhood of the ambiguous word. This information is captured in the model via the feature variables. The class variable on the other hand represents the various senses of a word or the semantic tags associated with it. The probability distribution is learned (estimated) from sense-tagged data, and is used to predict the most probable class (sense) for a given input.

This dissertation presents a novel, hybrid approach to learning probabilistic classifiers for WSD by combining an unsupervised learning technique, the Self Organizing Map (SOM) algorithm with Bayesian Learning (BL), a supervised learning technique. The SOM is used as an exploratory tool to automatically obtain a semantic landscape of Swahili[1]. This reveals the type of semantic classes (categories) that are deducible directly from data, and which would be used as a basis for a class-based disambiguation approach. The map also provides information regarding the most important linguistic cues necessary for semantic discrimination. The information obtained from the unsupervised learning step is incorporated into the design of Bayesian classifiers, where a classifier is constructed for each of the higher-level semantic categories. Disambiguation then reduces to a classification problem where semantic class membership is determined for a particular occurrence of the ambiguous word. The intended meaning is selected by choosing the English reading with a semantic class equivalence as the ambiguous word. WordNet tags are used to determine the semantic properties of the English words. This approach allows for WSD within a bilingual framework, without the need for parallel corpora, as is required by most other existing approaches.

[1] I refer to the Swahili language without its language-specific prefix *ki-*, following the widely accepted practice in Bantu linguistics.

The requirement of huge semantically-tagged training data has been described as a serious bottleneck facing the use of supervised learning methods for WSD. The methodology used in this dissertation alleviates the need for manually sense-tagged data by exploiting semantic similarity via distributional clustering to obtain annotated data from raw corpora. This is an important contribution especially for those languages that are deficient in computer-readable linguistic resources such as parallel corpora or semantic hierarchies such as the English WordNet. The method shows how resource-deficient languages can capitalize on resources available for other languages to facilitate development of their own resources and implement LT applications.

In total, Swahili Two-Level Parser (SWATWOL), a morphological parser for Swahili is used to pre-process and analyse Swahili texts obtained from the Helsinki corpus of Swahili. The resulting analyses are used in the creation of training data, based on carefully selected contextual features. The SOM algorithm is used to derive a semantic landscape of Swahili by clustering a set of unambiguous nouns occurring in the corpus. The resulting map is used to discover important semantic classes for Swahili and their corresponding discrimination cues. These are incorporated into the design of Bayesian word sense disambiguators. The Bayesian Modelling Toolbox (BMT) is used to learn the Bayesian classifiers from training data. The sense definitions for ambiguous words are obtained from the TUKI[2] Swahili-English dictionary, while WordNet, a computational lexicon for English provides the semantic link between ambiguous Swahili words and their English translations. The classifiers are tested on disambiguation and tagging tasks using different test sets comprising both ambiguous and unambiguous nouns. Standard evaluation measures are used for performance assessment.

[2] *Taasisi ya Uchunguzi ya Kiswahili* - Institute of Kiswahili Research, University of Dar es Salaam, Tanzania

1.1 Motivation and Research Objectives

The work undertaken in this study is done within the context of the Swahili Project which is headed by Professor Arvi Hurskainen of the University of Helsinki. Work on the development of computational tools for Swahili analysis started in 1985 with the design of a rudimentary morphological parsing program which has now evolved into a comprehensive language management system for Swahili. Development of a Swahili-English-Swahili MT system is one of the aims of the project, and this requires development of computational tools for comprehensive linguistic analysis of Swahili, from lexical and morphological analysis to syntax and semantic analysis. Work on the first phases of linguistic analysis has been successfully completed with the development of a morphological analyser and disambiguator. At the beginning of this study, the focus of research and development work was on syntax and semantic analysis, as the remaining linguistic analysis stages. The undertaken work focusses on semantic analysis and disambiguation for the MT system. In this context therefore, the main objectives of this study are to:

1. Perform a systematic analysis on the nature and extent of semantic ambiguity in Swahili with respect to English.

2. Develop a method for automatic WSD, also known as Target Word Selection (TWS) in the context of MT.

1.2 Swahili

Swahili is widely spoken in East Africa, where it serves as a lingua franca. It has approximately 80 million speakers spread across several countries such as Tanzania and Kenya, where it has an official status, Uganda where it is a national language, and in regions that border these countries in Malawi, Mozambique, the Democratic Republic of Congo, Rwanda, Ethiopia and Somalia.

Swahili is used in all spheres of daily life. In public life, it is used in political discourse, civil service, courts and the Tanzanian parliament. It is an important commercial language where it is widely used in marketing, business transactions and banking. It has a long tradition in music and the creative arts, resulting in a rich heritage in this area. The Swahili language has achieved the status of a language of scientific writing for its own academic community, as witnessed by the growing body of specialized publications in areas such as linguistics, literature and the social sciences. On the educational front, Swahili is taught at the primary and secondary school level and is also the medium of instruction in Tanzanian primary schools (Mulokozi 2002). It is also taught at the university level within Eastern Africa, and in several universities in other parts of Africa and across the globe. The growth and use of Swahili has been accelerated by its use in the media - numerous newspapers, radio and television broadcasts are produced in the language. The importance of Swahili as an African language designated for international communication on the African continent and beyond is evidenced by the numerous Swahili international broadcasts available such as the British Broadcasting Corporation, Voice of America, Deutsche Welle and Radio Japan, numerous on-line newspapers and most importantly, by its formal accreditation as a working language in Pan-African institutional settings such as the African Union.

Swahili is a Bantu language belonging to the Niger-Congo family. It is a highly inflecting language where both prefixed and suffixed morphemes play an important grammatical role. The functions of prefixes are particularly important in both nominal and verbal morphology. In the case of nouns, as is typical with Bantu languages, each noun belongs to a noun class which is signaled by a pair of prefixes attached to the nominal stem, denoting singular and plural forms. Verbs have an agglutinating structure where a system of affixes is used to mark various grammatical relations, such as subject, object, tense, aspect, and mood. There is a system of concordial agreement in which nouns must agree with the main verb of the sentence in class and number. Adjectives, possessive pronouns and demonstratives also agree in class and number with the noun they modify. Swahili has a fairly fixed word order (SVO) at the sentence level, where the subject precedes the verb and the object, while within constituent phrases, modifiers succeed the head. Therefore

adjectives, pronouns, determiners etc., follow the nouns they modify while adverbs come after the verb. For Swahili therefore, the complex morphological structure is a rich source of important syntactic and functional information, while grammatical relations can be differentiated through word ordering and indexing, providing useful cues for determining the semantic properties of words. The solution developed in this study exploits this linguistic information as detailed in chapter 3.

1.3 Dissertation Overview

The remainder of this dissertation is organized as follows:

Chapter 2 gives a basic introduction to WSD, which is the central concept in this study. In addition, a review of the main approaches to WSD that have been undertaken since the early 1950's to date, is presented. A brief discussion on systems that use cross-linguistic sense definitions as well as class-based WSD approaches is included, as these are comparable to this study. Also included is a brief discussion on the main Machine Learning (ML) methods for WSD.

In **Chapter 3**, the methodology employed in the development of the WSD solution is presented. The discussion covers a brief review of the resources, both linguistic and computational, that are required for solution development. The chapter is organized around three main themes:

i) A semantic exploratory phase using the SOM algorithm that reveals the important semantic distinctions (classes) for Swahili WSD that are directly inducible using overtly-marked linguistic features derived from textual data.

ii) An analysis of lexical ambiguity inherent between Swahili and English to identify the distinctions important for WSD, based on the classes identified in i).

iii) Design and training of Bayesian classifiers for WSD based on information obtained in i) and ii).

The performance of the WSD classifiers is evaluated in **Chapter 4**. Here, the performance metrics used in the evaluation are described, and the results presented. The learned classifiers are evaluated on disambiguation and tagging tasks using varying test sets to measure their performance. An analysis of the obtained results as well as the factors affecting disambiguation performance is presented.

Chapter 5 concludes the study by discussing the significance of the obtained results and recapitulates on the contributions and achievements made in the study. A discussion on the limitations of the work plus proposals for further work are presented.

Chapter 2
Related Work

The purpose of this chapter is to provide a brief review of the field of WSD. This is achieved by describing the major approaches to WSD that have been employed during the evolution of WSD research, from its inception in the 1950's to the current SENSEVAL[3] era (Kilgarriff 1998). A review of systems that define the WSD problem cross-lingually as well as those that employ a class-based strategy for WSD is also given, as these are particularly related to the approach employed in this dissertation. Where applicable, a brief discussion of ML techniques for WSD is included.

A comprehensive coverage of existing approaches is deemed important as it facilitates an understanding of the central problems in WSD research, and also provides a basis for comparing the solutions to these core problems, as implemented in this study. Therefore, in describing the main WSD approaches, particular attention is paid to the type of disambiguation information used, the required resources, system coverage and scalability, and to the granularity and representation of word senses, employed by each of these approaches.

The first section of this chapter, 2.1, gives a general introduction[4] to the WSD problem. Section 2.2 presents a timeline in WSD research, discussing the individual approaches whilst noting the factors that contributed to the progression from one WSD era to the next. Sections 2.3 and 2.4 review selected cross-lingual and class-based systems respectively.

---

[3] SENSEVAL is a series of workshops and competitions whose aim is to provide an evaluation framework for WSD systems. The strengths and weaknesses of various systems are evaluated on comparable tasks with respect to different words and varieties of languages.

[4] This section gives a very basic introduction to WSD with the sole objective of making this work accessible to a wider readership, in particular to researchers and linguists working with African languages, since the general field of LT and NLP in African languages is still largely in its infancy.

2.1 Word Sense Disambiguation

One of the first problems encountered by any NLP system is that of ambiguity. Ambiguity expresses itself at different levels. It could be at the part-of speech level where a lexical item can take one of several grammatical roles in a sentence. For example, the Swahili word '*kaa*' can be used as a verb to mean *dwell or sit* or as a noun to mean *a crab*, *charcoal* or *ember*. Another type of ambiguity common to NLP is that of prepositional attachment, where a prepositional phrase can attach to one of several constituents, yielding different parses for a sentence, and consequently, several possible meanings. For example, the Swahili sentence "*mgonjwa alikunywa uji na maziwa*" can mean "*the patient drank porridge and milk*" or "*the patient drank porridge with milk*", due to the ambiguity of the preposition '*na' and, by, with, also etc.*.

Highly accurate part-of-speech taggers and syntactic parsers have been developed for most languages, successfully addressing these type of ambiguities. The most pervasive ambiguity facing NLP today remains that of lexical ambiguity, where a word can have two or more associated meanings, depending on the context of use. The Swahili word '*kaa*' is a good example of this, where the verb reading is ambiguous between *dwell, sit* and *stay,* while the noun reading is ambiguous between *charcoal, crab* and *ember*. To resolve this type of ambiguity, knowledge of the different meanings that can be associated with an ambiguous word as well as the typical contexts in which they occur is vital. WSD is the process by which contextual information is employed to resolve lexical ambiguity and determine the intended meaning (sense) of an ambiguous word.

The history of WSD research is as old as that of MT. As early as 1960, Bar-Hillel, a prominent figure in early work in MT noted both the importance of WSD to MT, as well as its difficulty. His sceptic view on the ability of a machine to perform disambiguation of word senses was clearly evident when he famously proclaimed that "sense ambiguity could not be resolved by electronic computer either current or imaginable". He used the following example 2.1, containing the polysemous word *pen*, as evidence, arguing that even if *pen* were given only two senses, "writing implement" and "enclosure", the computer would

have no way of deciding between them (Bar-Hillel 1960) in (Ide & Véronis 1998).

Little John was looking for his toy box. (2.1)

Finally he found it.

The box was in the *pen.*

John was very happy.

One of the main reasons why WSD is so difficult is that meaning is generally vague in nature, and this makes it very difficult to define what the senses of a word actually are (Kilgarriff 1997). What constitutes a sense in natural language is the subject of serious debate, both in the fields of lexical semantics as well as computational linguistics. Many researchers have tried to ascertain the meanings of words by observing several examples of the contexts in which a word occurs, based on the hypothesis that a particular sense will typically occur in certain well-defined contexts. The problem with this approach is that a word can be used in very many different contexts, with some contexts representing only slightly varying meanings of the word, such that it becomes hard to characterise which are unique senses and which are not. This was clearly observed by Kelly and Stone (1975) when they stated that "the set of contexts in which a word appears with varying shades of meaning is not simply large, but indefinitely large...". For purposes of WSD, most researchers resort to using pre-defined sets of meanings as listed in standard dictionaries, rather than delving into analysing theories of defining meaning and senses. Most work on disambiguation has focussed on monolingual definitions of meaning following the work of lexical semanticists such as Cruse (1986), Levin (1993) and Pustejovsky (1995), who seek to quantify meaning dimensions within a single language. An alternative approach has been to use cross-linguistic correspondences for characterizing word meanings in language, where quantification of a word into senses depends on whether each sense can be uniquely translated in another language or set of languages. Examples of work following this line of cross-linguistic meaning quantification include Resnik & Yarowsky (1999), Ide (2000) and Gonzalo et al. (2002). The work presented in this dissertation focuses on

WSD within a MT context. As such, senses of words are expressed cross-lingually and the definition of these senses is obtained from the TUKI Swahili-English dictionary.

The other main reason that makes WSD such a difficult problem has to do with the different types of knowledge or information sources required for disambiguation. On closer analysis of the example given by Bar-Hillel in example 2.1, it is clear to see that this is a situation where selectional restrictions fail to disambiguate the word *pen*, since both senses indicate physical objects in which things may be placed, as indicated by the preposition *in* which applies to both of them. In this case, disambiguation is only possible if real world-knowledge regarding the relative sizes and uses of *pen* in the different senses is available. Also required are inferencing mechanisms that would make use of this knowledge and infer the intended sense of *pen* in the given example. Humans, in addition to making use of world knowledge for disambiguation, also use discourse or pragmatic information, where knowing the speaker's or writer's intentions can help one to resolve ambiguity. Unfortunately, formalizing all this information and rendering it in a form that is readily-usable by a computer has still not been accomplished. It is this ultimate dependence on world knowledge that has led WSD to be classified as an AI-complete[5] problem.

However, despite the seemingly insurmountable challenges facing WSD, success has been reported by various researchers, employing a broad range of disambiguation methods. There are those who concentrate on building knowledge-bases to capture real world knowledge and provide inferencing mechanisms that enable the computer to reason about the world, and thereby perform sense disambiguation. The major drawback associated with these approaches is the expense associated with manually creating knowledge bases. Consequently, the knowledge bases are small and the resulting disambiguation systems can only handle a handful of words from a simplified domain. On the other end of the spectrum are those researchers who choose to describe natural language using statistical methods, rather than try to explain it, as their knowledge-based counterparts do. Recent work in WSD has focussed on statistical methods and this has been influenced largely by the availability of huge electronic corpora as well as corresponding development of statistical tech-

[5] An AI-complete problem is one whose solution requires a solution to the general AI problems of reasoning about world knowledge.

niques for textual data mining. These techniques have been applied successfully to other tasks in NLP such as part-of-speech tagging and syntactic parsing, leading researchers to believe that they could also be successfully applied to the task of WSD. While statistical techniques have received criticism due to their lack of deep linguistic processing or understanding of natural language, they still offer various advantages that outweigh those offered by their more traditional knowledge-based counterparts. In addition to formally characterizing the uncertainty associated with word meanings, these methods also provide automatic or semi-automatic means of linguistic knowledge acquisition via data mining, and as a result, benefit from the concrete insights gained from a data-driven exploration of natural language (Lagus & Airola 2001, Bruce 1995).

Despite its associated difficulty, WSD is central to the success of most other LT applications. It has been identified as an important intermediate task that could significantly improve results of applications such as MT, IR, document classification, speech recognition, part-of-speech tagging, morphological and syntactic parsing. For MT, WSD is important when it comes to selecting the appropriate target language word for an ambiguous source language word. For example, to translate the Swahili noun '*kaa*' into one of its English equivalents: *crab, charcoal* or *ember*, a disambiguation algorithm that uses contextual evidence derived from the Swahili sentence would be necessary to determine which of these three senses is intended, and consequently make a translation decision. For IR, sense disambiguation would prevent the retrieval of irrelevant documents that contain query words of a different sense, while use of semantic tags could help in solving the prepositional phrase attachment problem.

2.2 Approaches to Word Sense Disambiguation

Most disambiguation approaches tend to focus on the identification of word-specific contextual indicators that can be used to distinguish between a word's senses. Efforts to acquire these clues or indicators have been characterized by their need for intensive human involvement for each word, which creates the associated problem of limited vocabulary coverage.

This is termed as the knowledge acquisition bottleneck in WSD literature. WSD systems can thus be classified based on how they attempt to deal with the knowledge acquisition bottleneck, by considering how they acquire disambiguation information. Using this criterion, a WSD system can be classified as knowledge-based, corpus-based or hybrid, and each of these approaches is briefly discussed in the following sub-sections. See Ide & Véronis (1998) for a detailed review.

2.2.1 Knowledge-based Approaches

Knowledge-based approaches encompass systems that rely on information from an explicit lexicon such as Machine Readable Dictionaries (MRD), thesauri, computational lexicons such as WordNet or (hand-crafted) knowledge bases.

Early Machine Translation Approaches

Much of the early work in WSD was carried out within the context of MT in the 1950's. The earliest approach was by Weaver (1949), where he argued the need for WSD in MT, as described in his memorandum. He introduced the notion of using a context window of size N from the neighbourhood of the ambiguous word, for WSD. He also realized and noted the important relationship between domain specificity and reduced word sense ambiguity, where the possible senses of a polysemous word are bound by the domain of use. Kaplan (1955) carried out experiments to determine the minimal size of N that is sufficient for WSD. He concluded that $N = 2$ was sufficient for WSD in most cases and that there was no significant improvement in WSD accuracy, when a bigger value of N or the entire sentence was used to provide contextual information. Masterman (1957) used *ROGET's* thesaurus to determine Latin-English translations based on the most frequently referred to thesaurus categories in a Latin sentence. His work laid the foundation for the use of statistical techniques for NLP.

As shown here, much of the foundation of WSD was laid in this period, but due to the lack of resources, both linguistic and computational, most of the ideas were not seriously tested. In subsequent years, most of these ideas have been tested and confirmed by

various researchers. For example, Gale et al. (1992*c*) and Yarowsky's (1995) 'one sense per discourse' echo Weaver's work on domain specificity of word senses, while several experiments on context window size confirm Kaplan's conclusions even for different languages e.g. Koutsoudas & Korfhage (1956) on Russian, and Choueka & Lusignan (1985) on French.

AI-Based Approaches

In the 1960's and 1970's, there was a lot of growth in AI research, and consequently, most of the methods that tackled WSD during this period used AI approaches. These systems relied on a wealth of both language and world knowledge, to determine the meaning of a word in context. Majority of these systems were grounded in language understanding theories and attempted to model deep knowledge of linguistic theory, especially in the area of syntax and semantics. Consequently, these systems tried to produce a semantic representation for an entire sentence in an attempt to capture its meaning, and from which word ambiguity problems would be solved. However, due to the pervasive nature of both structural and lexical ambiguity in natural language, a sentence can have several possible interpretations. In order to determine the correct interpretation, these systems adopted a strategy of combining syntactic, semantic and world knowledge and enforcement of constraint satisfaction, to produce syntactic and semantic representation of an entire sentence.

The scheme adopted for world knowledge representation as well as the process used to integrate syntactic, semantic and world knowledge, serve as the main distinguishing factors amongst these systems. Quillian (1961) used semantic networks[6] to represent world knowledge while Cottrell (1985), Waltz & Pollack (1985) and Eizirik et al. (1993) included syntactic information into the network as well. Other systems such as Hayes (1977) and Hirst (1987) used Frames[7], while Wilks's (1975) and Boguraev's (1979) case-based systems employed preference semantics to specify selectional restrictions for combinations of

[6] The nodes of the network are semantic representations of words or concepts, while the arcs represent relationships between concepts. Identification of word-sense associations is done through a process referred to as spreading activation.

[7] A frame represents a word as an entity and explicitly specifies its roles and relations to all the other words in the sentence.

lexical items in a sentence. These restrictions were used to determine which senses should be preferred over others, for a given context.

By the late 1980's, AI-based methods began to lose their appeal, largely due to the intensive manual labour that was required to create the knowledge bases. As a result, only relatively small knowledge-bases were created. This had the adverse effect of limiting most research work to 'toy' systems that had restrictions on the number of words, senses and syntactic constructs that could undergo analysis and disambiguation. Also, their insistence on deep syntactic and semantic analysis at the sentence level compounded the WSD problem, especially with hindsight of how difficult it is even today to obtain deep syntactic analysis of a sentence, let alone semantic analysis.

Dictionary-based Approaches

In the 1980's, there was a surge in computing machinery and a corresponding increase in the availability of electronic linguistic resources, popularly known as MRDs, as most publishers started to produce electronic versions of their products. This precipitated the shift from AI-based systems to the emergence of dictionary-based approaches. MRDs presented a viable solution to the knowledge acquisition bottleneck facing AI-based approaches since they provided comprehensive lexical coverage of natural language. This meant that systems no longer suffered vocabulary limitations, spurring interest in language processing of unrestricted text.

One of the first attempts to utilize these resources for WSD was Lesk (1986). His work was based on the observation that the coherence of a sentence is dependent on the cohesion of the words in it, meaning that the choice of one sense in a text is a function of the senses of the words close to it. He devised an algorithm that chooses the correct sense of a word by calculating the word overlap between the context sentence and the dictionary definition of the word in question. Lesk's work influenced most of the subsequent work in knowledge-based WSD such as McDonald et al. (1990), Véronis & Ide (1990), Wilks et al. (1990), Guthrie et al. (1991) and Cowie et al. (1992). Other machine readable resources that have been used in knowledge-based WSD include thesauri such as *ROGET's*

thesaurus that has been used severally by different researchers including Masterman (1957) and Yarowsky (1992), and lexicons such as *CyC*, *ACQUILEX*, *COMLEX*, *CORELEX* and WordNet Fellbaum (1998) in (Resnik 1999).

A major hindrance to dictionary-based techniques such as those based on Lesk's idea is their crucial dependence on similarity in wording between a text and the MRD. Dictionary definitions are usually too short to generate an overlap from which an adequate set of indicators can be obtained. Also, despite their well-structured information and increased vocabulary coverage, pre-coded knowledge sources suffer from limitations in domain-specific coverage and in coping with the introduction of new words.

2.2.2 Corpus-based Approaches

Corpus-based methods provide an alternative strategy for overcoming the lexical acquisition bottleneck, by obtaining information necessary for WSD directly from textual data. WSD is performed using information obtained by training statistical language models on a corpus. As noted in the preceding section, a major limitation of knowledge-based WSD systems is their reliance on pre-coded knowledge sources, which affects their inability to handle large vocabulary in a wide variety of contexts due to the associated expense of manual acquisition of lexical and disambiguation information. In an effort to overcome this problem, fuelled by the increased availability of natural language data in electronic form, WSD researchers have recently turned to corpora to help extend the coverage of existing systems as well as bootstrap or train new systems. These approaches have also benefitted from corresponding research in ML and statistical techniques, and especially, in their application to corpora, making it possible to obtain disambiguation information from textual data automatically. In addition, the success with which statistical techniques have been applied to other NLP tasks such as speech recognition, parsing and part-of-speech tagging has raised optimism that they can also be used for WSD work. In keeping with the latest trends in WSD research, this study adopts a corpus-based approach which offers the most promising solution to the knowledge acquisition bottleneck, by exploiting statistical learning techniques applied to corpora.

The earliest large-scale corpus-based approach to word meaning disambiguation was by Kelly & Stone (1975), who were working with a corpus of over 500,000 words. They sought to establish a set of word meanings perceived as useful for content analysis work. They manually developed an ordered set of disambiguation rules for each sense that was to be defined. These rules utilized a wide range of contextual features drawn from a ± 4 word window and included target word morphology as well as the identity, syntactic and semantic category of contextual words. Most subsequent work has focussed on the use of ML algorithms for the automatic acquisition and subsequent use of such contextual information for disambiguation.

Learning algorithms are categorized as statistical[8] or symbolic, where unlike statistical techniques, symbolic methods do not use probabilities explicitly. Examples of statistical learning techniques include Hidden Markov Models, log-linear models and BL, while symbolic methods include a wide array of algorithms such as decision trees, decision lists, transformation-based error-driven learning, instance-based learning, inductive logic programming, neural networks, genetic algorithms, clustering and support vector machines. Màrquez (2000) gives a detailed review of these methods and their application to various NLP tasks, including WSD. In ML, a distinction is usually made between supervised and unsupervised learning, see Mitchell (1997). In supervised learning, a set of *a priori* potential classes (senses in the case of WSD) are established before the learning process, while unsupervised learning means that the set of senses for a word are inferred *a posteriori* from text. However, as has been noted by Rigau et al. (1997), in the field of statistical NLP, unsupervised learning has also been used to mean an algorithm which does not require annotated training data, while those systems which require annotated training data are classified as supervised learning algorithms. Many corpus-based systems have been developed, and these encode disambiguation information using a broad range of contextual features such as collocations, co-occurrence information, syntax, case roles constraints etc. in different combinations. Using Rigau's definition for supervised vs. unsupervised systems, and considering the type of resources used to provide disambiguation information, Brown et al.

[8] Also referred to as probabilistic or stochastic

(1991*b*) and Ng & Lee (1996) are examples of supervised systems while Yarowsky (1992), Dagan & Itai (1994), Yarowsky (1995) and Pedersen & Bruce (1997) represent unsupervised WSD systems. Generally, better results in disambiguation have been achieved using supervised approaches (Màrquez 2000). BL, a supervised probabilistic method is selected for this study, and a detailed description of its theoretical foundations is given in section 3.5.2.

Despite the obvious benefits that corpus-based systems provide, they are also faced with certain setbacks and challenges. Although supervised systems have been purported to facilitate large-scale WSD, the requirement for annotated corpora has been a major setback to these systems, and as a result, most studies of this type have been limited to small sets of ambiguous words, usually less than twenty. To date, most annotated corpora have been prepared manually and this has limited the availability of such corpora. Research in the area of automatic annotation of texts or development of systems that exploit other resources with the aim of bypassing the requirement of annotated corpora, continues to gain considerable interest. This study is faced with this problem since there are no annotated corpora for Swahili, and proffers a solution which exploits distributional semantic properties of nouns using an unsupervised learning technique as described in detail in chapter 3, to automatically acquire annotated training data. The other challenging problem facing corpus-based approaches is that of data sparseness, which is characterized by disparity in the frequencies of word senses, where some senses do not occur at all in a given corpus, or occur very infrequently to be statistically significant. This poses problems for the ML algorithm since it will not learn how to accurately distinguish and disambiguate some senses. Again, this study takes a class-based approach to WSD as a solution to the data sparseness problem (see section 2.4). This shifts the sense disambiguation task from the word level, to a broader class level. In so doing, the data sparseness problem is addressed since training data is now not collected for a single word, but from several words that belong to a given class. This increases the probability of observing several occurrences in the available corpus, that are representative in meaning, to all the different senses for a given ambiguous

word. By coping relatively well with data sparsity (see section 3.5.2), BL complements the class-based approach well, in dealing with this problem.

As noted earlier, the successful application of statistical techniques to other NLP tasks raised optimism that these techniques could be applied to WSD. However, it is worth noting that the WSD problem is inherently much more difficult than say speech recognition or part-of-speech tagging. This is mainly due to the difficult problem of defining just what constitutes a sense of a word, and consequently determining how many senses a word has. For example, speech recognizers for English are trained to recognize approximately 625 triphone contexts, and as reported by Rabiner & Juang (1993), this task can be achieved with greater than 95% accuracy. Likewise, Brill et al. (1990), report 97% accuracy for a part-of-speech tagger trained on a corpus of 1.5 million words and a set of 64 part of speech tags. In contrast, a sense tagger based on a simple English learner's dictionary with about 55,000 words would have a tag set of 74,000 senses (Wilks et al. 1990). Similarly, a Swahili sense disambiguator for the approximately 3,000 ambiguous words listed in the TUKI dictionary would have to contend with approximately 10,000 senses. This means that the disambiguator would have to learn thousands of disambiguation rules to adequately disambiguate all the ambiguous words. This requires a considerably much larger corpus than would be required for say, a part-of-speech tagger or speech recognizer. It also implies that it would be beneficial to use abstract and generalized relations in constructing disambiguation rules, in order to make the WSD problem feasible, given the existing limitations associated to annotated corpora availability and sense distribution.

2.2.3 Hybrid Approaches

These approaches can neither be properly classified as knowledge or corpus-based, since they obtain disambiguation information from both corpora and explicit knowledge-bases. Luk's (1995) system is an example of a hybrid approach that combines information in MRD definitions with statistical information obtained from raw corpora. He uses textual definitions of senses from the LDOCE[9] to identify relations between senses. To determine

[9] Longman Dictionary Of Contemporary English

which of these relations are most useful for WSD, he uses a corpus to compute Mutual Information (MI) scores between these related senses.

Bootstrapping approaches where initial (seed) data comes from an explicit knowledge source which is then augmented with information derived from corpora, are another example of hybrid systems. Yarowsky's (1995) unsupervised system is a good example of a bootstrapping approach. He defines a small number of seed definitions for each of the senses of a word (the seeds can also be derived from dictionary definitions or lexicons such as WordNet synsets). He then uses the seed definitions to classify the 'obvious' cases in a corpus. Decision lists are used to learn generalizations based on the corpus instances that have already been classified. This process is repeated iteratively to the corpus, classifying more instances. Learning proceeds in this way until all corpus instances of the ambiguous word have been classified.

Hybrid systems aim to harness the strengths of the individual approaches while at the same time, overcoming specific limitations associated with a particular approach, to improve WSD accuracy. They operate on a 'knowledge-driven, corpus-supported' theme, utilizing as much information as possible from different sources. For example, Luk successfully exploits a lexical resource to reduce the amount of training data required for WSD, while Yarowsky's seeds provide initial knowledge, critical to the statistical learning phase.

2.3 Cross-lingual Word Sense Disambiguation

In this section, a review is given of approaches that have defined the WSD problem within a cross-lingual framework. In these systems, sense distinctions of an ambiguous word in one language are determined from its translation into another language or set of languages. This approach lends itself naturally to specific NLP applications such as MT and Cross-lingual Information Retrieval (CLIR) which, necessarily, involve two or more languages and hence demand a cross-lingual setting. More recently however, in an attempt to provide an alternative solution to the elusive philosophical and linguistic question as to what constitutes a word sense, some researchers have proposed that cross-lingual sense comparison

can be useful for sense disambiguation. This has served as a basis for some recent work in WSD such as Resnik & Yarowsky (1997), Ide (2000) and Diab & Resnik (2002) to name a few.

One of the earliest cross-lingual WSD studies is by Brown et al. (1991*a*). In this study, their aim is to investigate whether the addition of a WSD module to their statistical MT system (Brown et al. 1990), would have any impact on the MT results. Their system uses the English-French language pair and requires a word-aligned parallel corpus as well as part-of-speech taggers for both languages. Two-way ambiguity per word, for both languages is also assumed. The disambiguation process starts by extracting a set of the most frequent words for both languages. Each of these words is then described using a number of contextual features which capture information relating to the tense, part-of-speech, identity and position of contextual words, with respect to the ambiguous word. Different features are used for the two languages such as *tense-of-current-word, word-to-left, word-to-right, two-words-to-left, first-noun-to-left* etc. The Flip-flop algorithm (Nadas 1983) is used in conjunction with the splitting theorem (Breiman et al. 1984) based on MI[10], to make binary decisions between the different contextual features and the translations of the word in question. The translation of an English word is determined as that with the maximal MI with the French word. This method is evaluated *in-vivo* on 100 randomly-chosen English-French sentence pairs with the authors reporting an MT improvement of 8 percentage points, from 37% to 45%.

Another bilingual WSD approach is presented by Gale et al. (1992*c*). They also propose a solution to the knowledge-acquisition bottleneck by exploiting parallel corpora to obtain both training and testing materials. They do this by using translations as labels to annotate a set of polysemous words in a source language of a parallel corpus. This way, annotated corpora is created automatically. Their system uses the English-French language pair and is based on the Canadian Hansards. This system requires a parallel corpus that has been aligned at both the sentence and word level, from which training material is created. The disambiguation algorithm is supervised and consists of a training and a testing phase.

[10] A statistical measure of significance.

During the training phase, the sense of an occurrence of a polysemous word for a given context is identified. This is achieved by determining its translation, via its alignment to a target language word.

$$score(c) = \prod_{token\_in\_c} \frac{\Pr(token \mid sense_1)}{\Pr(token \mid sense_2)} \tag{2.2}$$

Equation (2.2) is used to obtain the context score of an occurrence of an ambiguous word. This equation is a variation on IR techniques where documents have been replaced with contexts. The context score is obtained by calculating the probability of a token appearing within a window of 50 tokens on either side of the ambiguous word. During the testing phase, the test instances of a polysemous word are identified and scored using equation (2.2). The correct sense is then selected on the basis of context score proximity by comparing the test scores with the training scores. This model ignores word order and collocational information when considering contextual information. Also, a smoothing approach that uses weighting is adopted to avoid the problems associated with sparse local token probabilities. The method is evaluated on six polysemous nouns, each having two distinct senses. These nouns, which have been chosen because they translate into distinct French words, are *duty, drug, land, language, position* and *sentence*. The *accuracy* score is used to measure the algorithm's performance and the authors report 90% overall accuracy for the six words.

Dagan & Itai (1994) present a new approach to WSD in one language by using statistical data from a monolingual corpus in another language. Their method focusses on the problem of TWS in MT. The resources required by this method include a target language monolingual corpus, a bilingual lexicon and parsers for both languages. It is evaluated on two language pairs, German-English and Hebrew-English, and imposes no restrictions on the number of senses per word. The disambiguation process begins by parsing the source language into syntactic tuples. They use a form of dependency parsing using SLOT grammars (McCord 1990) that identify syntactic relations such as *verb-subject, word-adjunct* etc. The next step involves identifying ambiguous words in the source language. In the context of this method, a source word is deemed ambiguous if there exist multiple transla-

tions for it in a bilingual lexicon and if it fits the frame of the specific source word instance in the source corpus. This definition allows for the pruning of alternative source senses on syntactic grounds. This results in a reduction of the ambiguities that the system has to deal with. The source syntactic tuples are then mapped to those of the target language. This is done by using the bilingual lexicon to translate the words in the source tuples. During translation, hand-coded rules are employed to handle cases where there are cross-lingual syntactic divergences. The final step is that of choosing the most appropriate translation tuple from the target language corpus, using a combination of filters. These filters are based on the occurrence frequency of the said tuple in the target language, a probabilistic model that determines the most probable target language tuple and a constraint propagation algorithm that handles ambiguities arising from multiple syntactic tuples in a sentence. The system is evaluated on randomly-selected examples comprising 103 ambiguous Hebrew words and 54 ambiguous German words. The authors report 68% applicability for Hebrew and 50% for German, where *applicability* is a coverage measure that determines how many cases are attempted out of all possible cases. *Precision*, which is a metric of how many found items in those retrieved are correct, is also used for performance evaluation, with Hebrew recording 91% and German 78%. This is against a Most Frequent Sense (MFS) baseline precision of 63% for Hebrew and 56% German, at the same applicability level. The lower performance on the German words is attributed to the change in corpus genre from the source test set to target language corpus genre.

Kikui's (1999) unsupervised approach is one of the more recent works in cross-lingual WSD, and he focuses on TWS for an English-Japanese MT system. Resource requirements include a bilingual dictionary and bilingual comparable corpora. For this study, Kikui uses 1994 newspaper articles of the New York Times and the Japanese Shinbon newspaper. The system does not place any restrictions on the degree of polysemy for ambiguous words. The disambiguation algorithm incorporates two unsupervised modules: The first algorithm is the distributional sense clustering algorithm which is based on Schütze's (1998) distributional clustering, and is used to obtain sense clusters for both corpora. It is first applied to the most frequent terms in the source language corpus, thereby

creating source language sense clusters. A source term list is then created using IR techniques to prune out infrequent words. Next, the bilingual dictionary is used to translate the term list, creating translation candidates. The distributional sense clustering algorithm is then applied to the target language, yielding target language sense clusters. The final step in TWS is achieved by applying the cosine similarity measure to the target language sense clusters and translation candidates, with those having the highest similarity values being selected as the correct translations. This method is evaluated on 120 test instances and achieves 79.1% accuracy against a manually-created gold standard.

Given the TWS task of the Swahili WSD system, senses of ambiguous words are taken to be their translations into English, like in the described systems. However, unlike the surveyed systems which rely on numerous resources such as bilingual comparable and/or aligned corpora, part-of-speech taggers and syntax parsers for both languages, this study seeks to develop a WSD solution that does not make extensive resource demands for both languages. Instead it relies only on monolingual corpora and parsers, exploiting existing computational and linguistic resources such as WordNet to provide the necessary semantic bridge between the two languages. This is seen as a vital step that extends NLP to (computational) resource deficient languages such as Swahili.

2.4 Class-based Word Sense Disambiguation

Corpus-based approaches rely on statistical data to estimate language models for different NLP tasks. Due to data sparseness in natural language, a major problem facing statistical NLP techniques is that of estimating the probabilities of events (e.g. co-occurrence relations, senses etc.) that were not observed in the training corpus. Class-based methods, which allow for the estimation of generalized class parameters as opposed to parameters for individual words, have been adopted as one approach to solving the sparse data problem.

Yarowsky (1992) presents an approach to WSD that uses classes of words to derive models that can be used to disambiguate individual words in context. He uses Grolier's encyclopedia to learn statistical models of the major *ROGET's* thesaurus categories which

serve as approximations of conceptual classes. These categories correspond to sense distinctions and sense disambiguation thus involves selecting the most likely category for a word in context. The procedure for learning the category models starts with the collection of representative contexts for each of the 1042 *ROGET's* categories, from a ± 50 word window. The next step involves identifying salient words from the collective contexts that are highly indicative of a particular category.

$$\frac{\Pr(w|RCat)}{\Pr(w)} \tag{2.3}$$

Using equation (2.3), an estimate much like the MI metric is computed. This is an estimate of the probability of a word w, appearing in the context of a *ROGET* category ($RCat$), divided by its overall probability in the training corpus. The log of the salience is used as the individual word's weight in the statistical model of that category. The disambiguation step uses the resulting weights to predict the appropriate category for a polysemous word appearing in a new context. The algorithm achieved an overall accuracy of 92% when tested on 12 polysemous words averaging three sense distinctions.

Resnik (1999) presents an algorithm that disambiguates noun groupings, as opposed to individual words, with respect to WordNet senses. He assumes the existence of noun groupings that have been obtained via some black-box procedure, and whose relatedness has been established. In his experiments, he uses groupings derived from distributional clustering experiments (Brown et al. 1992, Schütze 1993) and thesaurus classes such as *ROGETs* and Grefenstette's (1994) Machine-generated thesaurus. He then devises an algorithm to determine which WordNet sense (class) subsumes all the members of the group. With this, the group is disambiguated with respect to WordNet's IS-A[11] hierarchy. He computes semantic similarity for all the group members using equation (2.4), to determine the concept (WordNet sense) that is the most informative subsumer (closest common ancestor) for all the group nouns. The semantic similarity of two words $W_1$and W_2 is calculated as:

$$sim(W_1,W_2) = \max_{c \in subsumers(w_1,w_2)} [-\log \Pr(c)] \tag{2.4}$$

[11] A hierarchy of subtype/supertype relationships.

where subsumers(W_1,W_2) is the set of WordNet synsets that subsume (i.e. are ancestors of) both W_1 and W_2, in any sense of either word. His algorithm is based on the observation that when two or more polysemous words are similar, their most informative subsumer provides information about which sense of each word is the most relevant. The algorithm was tested on 125 test cases and achieved an accuracy of 60% against a random-selection baseline of 34% and an upper bound of 67% set by two human judges, on average.

This study adopts a class-based approach much along Yarowsky's model, but exploits distributional clustering of Swahili nouns to automatically determine the semantic classes that form the basis of the WSD system, rather than relying on external definitions for such classes such as ROGET's thesaurus, which are unavailable for Swahili and most other less-studied languages. Furthermore, such a data-driven approach is preferable since the WSD solution is consequently based on semantic classes whose distinguishing properties have been ascertained to be available from textual data, and whose coverage is thus expected to be very high, approaching 100%.

2.5 Summary

A basic introduction to the field of WSD and statistical NLP has been presented in this chapter. A survey of the major approaches to WSD has been presented, emphasizing the key WSD research problems that should be addressed by any type of solution. The knowledge acquisition bottleneck has been singled out as a major challenge for WSD, and whose solution influences directly or indirectly, the training methodology adopted, the information and resources required for disambiguation as well as the coverage and scalability of the developed system. This study pursues a corpus-based methodology where ML techniques applied to a monolingual corpus of Swahili are employed in the training of the WSD system. The study adopts a class-based approach and uses the SOM algorithm to automatically determine what these classes should be. The classifiers for WSD are modelled as Bayesian networks which complement the class-based approach to get the most out of the available

training data. The methodology followed to realise the WSD solution is presented in detail in chapter 3.

Chapter 3
Word Sense Disambiguation using Bayesian Networks

3.1 Introduction

In this chapter, a hybrid WSD method for Swahili nouns is presented. It relies on ML techniques that acquire knowledge for disambiguation i.e. learn how to disambiguate, by combining information from a variety of sources - a corpus of Swahili, a Swahili-English dictionary and publicly available linguistic resources for English, namely WordNet, a computational lexicon for English and Levin's (1993) classification of English verbs.

A precise description of the problem under study and the methodology employed to solve it is presented in this chapter. In section 3.1.1, the problem definition is presented. This includes a discussion on the specifics of the WSD task under consideration such as the choice of sense tags, sense granularity, test and training data preparation and evaluation of system performance. A general overview of the WSD solution is also presented in this section. In section 3.2 a detailed description of the resource requirements, both linguistic and computational, is given. A detailed step by step description of the individual phases of the system is presented in subsequent sections 3.3 - 3.5. The discussion provides a brief review of the relevant theoretical background, implementation specifics as well as intermediate results where applicable. System evaluation is presented in chapter 4.

3.1.1 Problem Definition

The WSD problem is that of associating an occurrence of an ambiguous word with one of its senses. In order to do this, first, an inventory of the senses associated with each word to be disambiguated must be available; second, a mechanism to associate word senses in

context to individual senses must be developed, and thirdly, an evaluation procedure to measure how well this disambiguation mechanism performs must be adopted.

Important issues concerning the senses include determining the source of the senses such as dictionaries or usage contexts derived from corpora etc., and the level of sense granularity to be tackled. Another important decision that has to be made is the meaning representation scheme that will be used since this influences the design of the disambiguation algorithm. Designing the actual disambiguation mechanism involves the construction of disambiguation rules and their subsequent application to a real disambiguation problem, achieving WSD. The key issues to be considered here are the source of the disambiguation information, the construction of rules using this information and the criteria for selecting the correct sense of an ambiguous word in context, using these rules. Evaluation involves determining appropriate evaluation metrics, choosing test words and acquisition of test data, as well as obtaining a gold standard for evaluation.

Given that WSD is usually undertaken as an intermediate step for other NLP tasks, the application area for which the WSD solution is developed bears important consequences on various aspects of the WSD problem specification. In this dissertation, the problem of WSD is undertaken within the context of Swahili-English MT, and thus the WSD problem here is essentially TWS - choosing the most appropriate English translation for an ambiguous Swahili noun, in a given context. Given this background, the senses of a word are taken to be its English translations and the TUKI Swahili-English bilingual dictionary is used as the sense inventory. The granularity of the sense distinctions to be considered is determined empirically via a data exploratory phase which identifies the type of semantic distinctions that can be made, given the available linguistic information.

TWS requires that there exist a mechanism that associates the meaning representations for individual senses of a word to the equivalent target language translation. WordNet noun classes have been chosen for this purpose since all English nouns are already associated with a semantic (WordNet) tag, and the problem that remains is that of associating the context of an ambiguous Swahili noun with one of these tags. WSD is then achieved by selecting the English translation whose semantic tag matches the WordNet tag selected

for the given context. The disambiguation problem thus entails determining the semantic properties (WordNet tag) of an ambiguous Swahili noun in context.

To achieve disambiguation, the study employs corpus-based techniques, where a supervised ML algorithm, namely, BL is used to learn disambiguation rules automatically from a training corpus of Swahili where each example has been annotated using WordNet tags. The disambiguation rules are represented in the form of probabilistic classifiers which when given an occurrence of an ambiguous word as input, produce its semantic classification (WordNet tag). This tag is then used to select the appropriate English translation, via WordNet tag equivalence matching.

Evaluation of the WSD algorithm is done using standard performance evaluation metrics such as precision, recall, F_1 measure and accuracy. More on the specifics of evaluation are presented in chapter 4. In total, the methodology to be presented will be justified both in terms of its theoretical foundations and by the performance of classifiers developed per its specification.

WSD solution Overview

To ensure that the WSD solution suffices for its purpose, i.e. TWS for a Swahili-English MT system, it is instructive to determine the exact nature and extent of the ambiguity inherent between these two languages. This analysis provides a guideline to the semantic distinctions that the WSD system would have to be able to make in order to achieve high coverage and good performance.

Acquiring the knowledge required to do WSD has been highlighted as a serious challenge in the construction of WSD systems. As presented in chapter 2, various approaches have been adopted with an attempt to overcome this knowledge acquisition bottleneck. Corpus-based approaches that obtain disambiguation information automatically from textual data are the most promising in this respect. The WSD solution presented here is data-driven where determination and acquisition of useful information for WSD is done automatically. The emphasis here is on *data-driven* and *automatic* as this ensures that only that knowledge which is explicit in the language, and which is directly usable by a computer

system is used to make disambiguation decisions. This not only guarantees reliability and consistency in performance, but also renders the solution language-independent, meaning that it can be replicated for any language pair.

The main phases of solution development are:

1. **Obtaining a semantic landscape of Swahili Nouns**: The SOM algorithm is applied to data vectors derived from a corpus of Swahili, to obtain an initial grouping of nouns into clusters based on semantic similarity. This is an exploratory phase that is done to determine the semantic distinctions that can be made using linguistic features derived from text, and also obtain a mapping of WordNet noun classes to Swahili semantic classes. To do this, Model Based Clustering (MBC) is applied to the SOM codebook vectors as explained in section 3.3.2, with the aim of later refining the cluster boundaries for further analysis. For each cluster, its members and associated WordNet tags are analysed and together with the corresponding component maps, a set of semantic classes, each labelled with a unique name, is obtained from the clusters. These classes represent the ambiguities that the system can handle, while the class labels are used in the automatic annotating of training data, obtained from class members.

2. **Ambiguity analysis of Swahili**: In this step, a thorough analysis into the nature and extent of ambiguity in Swahili with respect to English is done. This is achieved using the Swahili-English bilingual dictionary, WordNet noun classes and the semantic classes obtained for Swahili nouns in step 1).

3. **Learning Bayesian Classifiers for WSD**: The semantic classes obtained in step 1) are used in the implementation of a class-based WSD solution for the ambiguities identified in 2), where the disambiguation task is essentially reduced to a classification problem. Choosing the most appropriate sense for a given occurrence of an ambiguous word is done by determining membership in one of the semantic classes, and selecting the English reading that exhibits/possesses the corresponding semantic feature(s). The probabilistic classifiers for each of the semantic classes are modelled as Bayesian

Belief Networks (BBN), with the latter being learned from the training data generated in step 1).

3.2 Resources

3.2.1 WordNet

WordNet (Fellbaum 1998) is a semantic lexicon for the English language developed and maintained at the cognitive science laboratory of Princeton University, New Jersey. Its design was inspired by current psycholinguistic theories of human lexical memory, and was developed under the direction of psychology professor George Miller.

WordNet divides the English lexicon into five categories: nouns, verbs, adjectives, adverbs and function words. The words are further organized into sets of synonyms referred to as synsets, each representing one underlying lexical concept e.g. {toddler, yearling, tot, bambino} represents a young child. Since WordNet organizes lexical information in terms of word meanings rather than word forms, semantic relations such as hypernymy, antonymy, hyponymy etc., are used to link the various synsets. WordNet further partitions words (based on their word category) into distinct hierarchies using a set of semantic primes or generic concepts. These hierarchies correspond to relatively distinct semantic fields but are not mutually exclusive. Nouns belong to one of 26 semantic types e.g. person, animal, event etc., while verbs are categorised into 15 different verb types e.g. motion, communication, consumption etc. As an example, figure 3.1 shows hyponymic relations among seven semantic components, denoting tangible (concrete) entities:

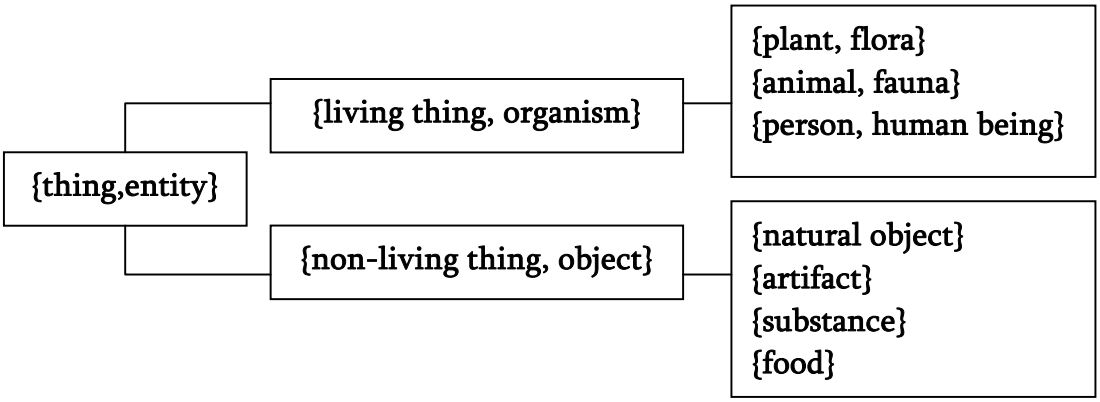

Figure 3.1: WordNet hyponymy relations for concrete entities

WordNet contains about 140,000 words organized into over 110,000 synsets, creating a comprehensive dictionary-thesaurus combination. WordNet's support for automatic text analysis and AI applications, coupled with its free online accessibility has contributed to its widespread usage as evidenced by the numerous NLP applications that use it as a semantic resource[12].

The WordNet database is critical to the WSD solution presented in this dissertation. WordNet tags are used to provide the semantic linkage between the WSD classifier's decision and the English translations of the ambiguous word, enabling TWS. WordNet verb tags are also used as predicate-argument contextual features as described in section 3.3.1.

3.2.2 Levin Verb Classes

Levin (1993) has organized 4183 verbs into 191 classes[13] according to a verb's behaviour with respect to certain syntactic alternations in the expression of arguments that affect meaning. The motivating principle is that verbal meaning determines syntactic realizations. Syntax therefore, serves as an important constraint on the possible meanings for

[12] A comprehensive WordNet bibliography is located at http://engr.smu.edu/~rada/wnb/

[13] The electronic version of Levin's verb classes is publicly available at http://www-personal.umich.edu/~jlawler/levin.html

a given verb by corroborating certain classes and disqualifying others. Figure 3.2 shows four classes for *contact* verbs, illustrating the distinctions that the various classes make. As shown, these verbs are categorized based on the type of alternations that they participate in.

Alternations for verbs of contact:

conative:

Jean moved the table.
*Jean moved at the table.

body-part possessor ascension:

Janet broke Bill's finger.
*Janet broke Bill on the finger.

middle construction:

Bread cuts easily.
*Cats touch easily.

Verb Classes

Alternation	Touch	Hit	Cut	Break
conative	N	Y	Y	N
body-part possessor ascension	Y	Y	Y	N
middle	N	N	Y	Y

Examples of verbs for each class

Touch:	kiss, sting, tickle
Hit:	bash, hammer, tap
Cut:	chip, hack, scratch
Break:	crack, split, tear

Figure 3.2: Examples of Levin's verb classes

Like WordNet, these verb classes have been used as predicate-argument contextual features that capture the grammatical relation between a verb and its dependents.

3.2.3 SOM Toolbox

The SOM Toolbox (Vesanto et al. 2000) is a public[14] domain function library for MATLAB 5 that implements the SOM algorithm. It has been developed at the Neural Networks Research Centre, Helsinki University of Technology. MATLAB by Mathworks Inc., provides an excellent environment for scientific computation and analysis. It employs a high-level

[14] SOM Toolbox Software and Documentation is available at http://www.cis.hut.fi/projects/somtoolbox/

programming language which makes it suitable for fast prototyping and customization, and has a strong support for graphics and visualization. The SOM Toolbox takes advantage of these strengths to provide an efficient, customizable and user-friendly implementation of the SOM algorithm. The toolbox can be used to preprocess data, initialize and train SOMs using a range of different topologies and visualizations. This enables a researcher to perform varied analyses of the properties of the SOMs and the data. A discussion of the SOM algorithm and its application to data categorization is presented in section 3.3.2.

3.2.4 Bayesian Modelling Toolbox

The BMT[15] is a data analysis tool for dependence and classification modelling developed by the Complex Systems Computation Group (CoSCo) at the Helsinki Institute for Information Technology. The tools enable analysis of data for multivariate probabilistic dependencies which are represented using Bayesian networks. The theoretical foundations of Bayesian modelling are discussed in detail in section 3.5. The specific theoretical design principles adopted in implementing the BMT are also included in the discussion. The toolbox has been used successfully for various applications such as Ruohotie et al. (2001). In this study, the toolbox is used to induce probabilistic classifiers for WSD from data and facilitate testing of the same via probabilistic inference.

3.2.5 SALAMA (Swahili Language Manager)

SALAMA is a suite of computational tools developed by Hurskainen at the University of Helsinki, for processing Swahili texts (Hurskainen 1999). Tools for linguistic analysis include a lemmatiser, morphological analyser & disambiguator and a syntactic mapper that performs surface syntax analysis, while end-user utilities include a spell-checker and hyphenator for Swahili.

Of particular importance to this study are the linguistic analysis tools and specifically the morphological analyser & disambiguator, SWATWOL and the shallow parser Swahili

[15] An online version of these tools, the B-Course service, is located at http://b-course.hiit.fi or http://b-course.cs.helsinki.fi and is freely available for educational and research purposes.

Constraint Grammar Parser (SWACGP). SWATWOL[16] has been so named as it is based on Koskenniemi's (1983) Two-level model for morphological analysis and generation. This Two-level model consists of a lexicon and rules (language-specific components) combined with a runtime engine applicable to all languages which make it language-independent. Hurskainen (1992) has developed the Swahili-specific components for SWATWOL which consist of the annotation scheme, lexicon and rules. The annotation scheme defines an extensive set of tags[17] used to code various linguistic properties of word forms such as morphology (part of speech, derivational and inflectional features), syntax, etymology, some semantic features and domain tags e.g. health care. The lexicon specifies the morphemes and words of the language that can be processed by SWATWOL. It was compiled from various word-lists, dictionaries and material obtained from the Swahili corpus and currently recognizes at least 45,000 words. The two-level rules specify the relation between the lexical and textual (surface) representations of words. They constrain the surface realization of lexical strings by specifying particular lexical/surface correspondences and the environment in which these correspondences are allowed, required or prohibited. SWACGP (Hurskainen 1996, Hurskainen 2004*b*) is a constraint-grammar parser which disambiguates ambiguous readings produced by SWATWOL. It also performs surface syntax tagging of word forms. It is based on the language-independent constraint grammar parser (Karlsson 1990, Karlsson et al. 1995, Tapanainen 1996) and a Swahili rule file that presently contains at least 1,200 rules prepared by Hurskainen. SWACGP achieves a good performance with a morphological ambiguity[18] residue of 8% for fiction/prose texts and 5% for newspaper texts (Hurskainen 1996:572). On average approximately 94% of ambiguous readings are successfully disambiguated. This figure improves even further to 97% when SWA-GUESS, a heuristic disambiguator is applied to the remaining ambiguities.

---

[16] In the body of the text, SWATWOL is used to refer to both the morphological parser and the shallow syntax parser SWACGP.

[17] The full tag set is located at http://www.aakkl.helsinki.fi/cameel/corpus/swatags.pdf

[18] The ambiguities handled at this level are morphological rather than semantic and include part of speech ambiguities e.g. noun vs adverb, adjective vs adverb, noun vs conjunction etc, genitive markers e.g. *ya* vs *wa* and possessive pronouns.

The first step in text analysis using SALAMA is that of pre-processing where the text is rendered ready for further linguistic analysis. This involves operations such as marking of sentence boundaries, separation of punctuation and diacritics from words, identifying multi-word terms and reduction of upper case to lower case whilst marking initial capitals. The pre-processed text is then analysed using SWATWOL and SWACGP, producing output which lists for each word form in the original text, a set of tags that describe its morphological and syntactic form. Also included is a list of English glosses for each word. For example, figure 3.3 shows SWATWOL's output for the sentence given in example 3.5.

Askari hao wamesema kuwa walipofika kwenye tukio, (3.5)

waliwazuia wafanyakazi na wapangaji wote wa jengo hilo kuingia

Those police have said that when they arrived at the scene,

they prevented all the employees and tenants of that building from entering

```
"<*askari>" "askari" N CAP 9/10-0-SG { soldier , guard } HUM
"<hao>" "hao" PRON DEM :hV ASS-OBJ 1/2-PL { these }
"<wamesema>" "sema" V 1/2-PL3-SP VFIN PERF:me { say , speak , scold , speak
against , advise , counsel , backbite , badmouth } SV SVO
"<kuwa>" "kuwa" CONJ **CLB { that }
"<walipofika>" "fika" V 1/2-PL3-SP VFIN PAST 16-SG-REL { arrive } SV
"<kwenye>" "kwenye" PREP { in , at , about }
"<tukio>" "tukio" N 5a/6-SG DER:verb (tukia) DER:io { event , happening ,
occurrence }
"<,>" "," COMMA
"<waliwazuia>" "zuia" V 1/2-PL3-SP VFIN PAST 1/2-PL3-OBJ OBJ { stop , restrain
, prevent , obstruct , support } SVO
"<wafanyakazi>" "mfanyakazi" N 1/2-PL DER:zi { worker , employee }
"<na>" "na" CC { and }
"<wapangaji>" "mpangaji" N 1/2-PL { arranger , filer } DER:ji
"<wote>" "wote" PRON :ote 1/2-PL { all }
"<wa>" "wa" GEN-CON 1/2-PL
"<jengo>" "jengo" N 5a/6-SG DER:verb (jenga) DER:o { building , construction }
"<hilo>" "hilo" PRON DEM :hV ASS-OBJ 5/6-SG { this }
"<kuingia>" "ingia" V INF { enter , get in , go into , incur , pierce , matriculate , join
a group/association/party } SV
"<.$>"
```

Figure 3.3: Morphological analysis and disambiguation output

The output from this morphological analysis and disambiguation stage forms the basis for different applications such as spell-checkers and hyphenators as well as Sewangi's (2001) domain-based terminology extraction from Swahili texts. The contextual information used for semantic disambiguation in this study has also been derived from this output.

3.2.6 Helsinki Corpus of Swahili (HCS)

The HCS[19] is an annotated corpus of standard Swahili texts that has been compiled at the Institute for Asian and African studies, University of Helsinki. The annotation has been done using SALAMA and contains the information described in section 3.2.5. The corpus is made up of a mixed genre of texts including religious texts (Bible, Qur´an), newspaper texts (both electronic and print), parliamentary proceedings from Tanzania and books containing prose text, fiction, educational and scientific materials. Currently the total size of the corpus is 12.5 million words though material is constantly being added to it.

The data used both for training and testing the WSD solution has been obtained from randomly selected texts from this corpus.

3.2.7 TUKI Swahili-English Dictionary

The TUKI Swahili-English dictionary is a standard dictionary of modern Swahili compiled at the Institute of Kiswahili Research at the University of Dar es Salaam, Tanzania. It claims to have more than 30,000 head words, but Hurskainen's (2004*a*) computational testing of Swahili dictionaries using SWATWOL reduces this number to 14, 533. The dictionary is available in electronic format as a simple text file, and had to be edited to produce listings of words by part of speech as well as lists of ambiguous and unambiguous words.

[19] Access is restricted to authorized users. Requests for authorization can be made at www.csc.fi/kielipankki/aineistot/hcs/index.phtml.en

3.3 Obtaining a Semantic Landscape of Swahili Nouns

For MT, the task of WSD is synonymous with TWS, where it suffices to distinguish amongst the different competing word senses of the source language word, such that the corresponding target language word can be selected as the right translation equivalent. Distinguishing senses from one another is a different task from that of defining their exact meanings. For the former, emphasis is on developing a criterion that is used to separate the different senses from each other, without caring to define what each sense 'means'. For example, for the ambiguous noun '*kaa*' with two possible translations *charcoal* and *crab*, a WSD algorithm can use the semantic property **ANIMATE** as a semantic distinction to determine if a particular occurrence of '*kaa*' refers to the animate reading *crab* or the inanimate one *charcoal*, without having to further define the meanings of *charcoal* or *crab*.

Therefore, to construct a WSD system that suffices for this sense discrimination task within a Swahili-English MT application, it is instructive to first and foremost determine the nature of semantic (lexical) ambiguity that is prevalent between the two languages, and by extension, the semantic distinctions that a WSD system for Swahili nouns should be capable of making. To accomplish this, a system of meaning representation is required to express the meaning of different word senses. This provides a framework for determining what types of ambiguities exist in the language pair under study. A system of semantic categorization is adopted where senses are associated with semantic categories (classes). Members of a given category share common semantic properties or attributes that distinguish them from those of a different category. The categories reflect the conceptual organization of the domain in which the WSD system must operate and should therefore represent the semantic properties that are necessary and sufficient for sense discrimination, given the identified ambiguity types.

The choice of which categories to include in a meaning representation scheme is a complicated one, but as Lenci (2001) notes, the specific application typically biases this choice to include those categories that allow for the organisation of the domain knowledge in a manner that is most needed for the given purpose. In this study, WordNet's 26 noun

Noun	Translation	WordNet Tag
Kichaa	Lunatic	Person
	Lunacy	State
	Bunch (of fruits)	Group
Bakora	Walking stick	Artifact
	Stroke	Act
	Apprenticeship fees	Possession
Mkoa	Province, Region	Location
	Metal bar	Artifact

Table 3.1: WordNet tagging of Swahili noun senses

classes (class hypernyms) are used to represent the meanings of individual senses (translations) of the ambiguous words, as shown in table 3.1.

As mentioned in section 3.2.1, WordNet was initially intended for psycholinguistic purposes, and even though it has been successfully used as a semantic dictionary in many NLP applications, the meaning distinctions it makes are often very subtle and fine-grained. This results in a large number of word senses which more often than not, are not very useful for many NLP tasks. Also, some of its noun class definitions reflect this psycholinguistic bias and may not be entirely compatible with a meaning representation scheme that relies only on linguistic attributes or behavior to determine semantic classes and their members, such as that adopted in this study.

For example, WordNet makes a distinction between natural objects (**object**) such as *rivers, mountains, hills* etc. and man-made objects (**artifact**) such as *pool, houses, tables, cars* etc. However, on examining the linguistic behavior of say a *pool* and a *river*, it is evident that they occur in similar linguistic contexts and function as locations i.e. places that people (animates) can be in/on, can go to, can swim in etc., and should necessarily be classified as **locations**. Another example are the **communication** and **cognition** tags that represent nouns denoting communicative and cognitive processes and contents respectively. Following this definition, concrete nouns such as *book* or *magazine* are classified together with abstract nouns such as *request, song* or *command*, since they all have something to do with communication.

For this reason, it was found necessary to define new categories by retagging or reorganizing WordNet classes to reflect the linguistic behavior of different semantic word types. In addition, an important consideration when choosing these new categories is that they should be distinguishable from each other using linguistic evidence derived from textual corpora. This requirement is critical since the WSD solution presented here relies on semantic class membership determination using only contextual linguistic evidence to decide what the 'meaning' of the ambiguous word in context is, and consequently select the appropriate translation. The disambiguation task has been structured as a classification problem where different senses of a word are associated with different semantic classes, and disambiguation therefore involves determining class membership.

To determine the semantic classes that are inducible using linguistic evidence derived from HCS and how WordNet's noun classes correspond to them, a semantic exploratory phase was carried out. The SOM was used to obtain a semantic clustering of unambiguous Swahili nouns. From these clusters, a minimal set of semantic classes sufficient for the WSD task and whose distinguishing properties can be automatically determined from Swahili textual data, were selected.

To use the SOM for this purpose, a set of linguistic features deemed important for semantic discrimination has to be selected and used to obtain training data for the SOM algorithm. Section 3.3.1 discusses the Swahili language with emphasis on its linguistic structure and selection of important contextual features. An overview of the SOM algorithm and its application in the determination of semantic categories is then presented in section 3.3.2.

3.3.1 Context Features

One of the most important tasks in ML is that of feature selection. This is the stage where the intrinsic domain knowledge is brought to the fore and incorporated into the system. In this study, knowledge of the linguistic structure, functions and interaction of various Swahili language elements is a pre-requisite to designing the WSD system. As explained in section 3.3, the success of the WSD system depends on identifying a compact set of

semantic classes whose distinguishing features or properties are reflected in the linguistic behaviour of words, and are thus obtainable from overtly-marked features in text. To achieve this goal, a set of contextual features with high predictive capability for different semantic classes has to be identified, ensuring that each feature can be easily extracted from the morphological tagset of the SWATWOL analyser. The features selected for the study are based on the linguistic properties highlighted in section 1.2. They are discussed below and have been organised according to the different knowledge types they represent (McRoy 1992, Agirre & Martinez 2001).

1. Morphological Features

   a) Noun prefix: There is a lot of discussion in the literature as to whether Swahili noun classes encode any semantic classification or not, with opinions ranging from yes to no but with the majority lying somewhere in between as Contini-Morava's (1997) discussion on the different positions on this issue shows. Some of the classes exhibit semantic consistency such as noun class 1/2 (denoting class 1 for singular and 2 for plural) which contains nouns that denote human beings, save for a few exceptions such as '*mdudu' insect* and '*mnyama' animal*, while others such as class 9/10 are a mixed bag of different semantic types such as humans, animals, artifacts etc. Nonetheless, the noun prefix is an important feature in the language which may have some semantic implications. Table 3.2 shows the different noun classes and their corresponding prefixes[20].

   b) Subject prefix: The subject prefix of a verb agrees with the subject noun and provides information about the subject, without even having to know what the actual subject noun is. This feature provides very important semantic information since there is an animate subject prefix *a-* associated with all[21] animate nouns, regardless of the noun class. In example 3.6, the noun prefix for the subject noun '*mtoto'* is *m-* (class 1) while that of example 3.7 '*madereva'* is *ma-* (class 6). However, they both take animate subject prefixes in the verb, *a-* (animate singular)

[20] 0 (zero) indicates a missing/absent noun class prefix.

[21] Diminutives and augmentatives are an exception to this as they take the *ki-/vi* prefixes.

Class	**Singular**	**Plural**
1/2	*m-toto* child	*wa-toto* children
3/4	*m-ti* tree	*mi-ti* trees
5/6	*ji-cho* eye	*ma-cho* eyes
5a/6	*0-somo* study	*ma-somo* studies
6	*ma-ji* water	*ma-ji* water
7/8	*ki-tabu* book	*vi-tabu* books
9/6	*0-dereva* driver	*ma-dereva* drivers
9/10	*0-taa* lamp *m-bwa* dog	*0-taa* lamps *m-bwa* dogs
11	*u-huru* freedom	*u-huru* freedom
11/6	*u-gonjwa* disease	*ma-gonjwa* diseases
11/10	*u-kuta* wall	*0-kuta* walls
15 (nominal infinitive	*kusoma* reading	
16-18 (locatives)	*ha-pa, hu-ku, hu-mu* here (within)	

Table 3.2: Swahili noun classes

and *wa-* (animate plural), making this feature a highly predictive indicator for animacy.

m-toto (SG-child) **a**-*na-enda* (3SG-PRES-go) *shule* (school) (3.6)

"The child is going to school"

ma-dereva (PL-driver) **wa**-*na-enda* (3PL-PRES-go) *kazi* (work) (3.7)

"The drivers are going to work"

c) Reflexive marker: The verbal infix *-ji-* expresses reflexivity, a property associated with animate (typically human) subjects or institutional nouns that can take on human properties e.g. *'bunge' parliament*, *'chama' meeting* etc. (example 3.8). From the reflexive infix, selectional preference information regarding the type of verbs that require animate subjects and that can take human objects is obtained. Thus, the verb can be subsequently used as an indicator of animate or institutional nouns.

Juma (Juma) **a**-*li-ji-kata* (3SG-PAST-REFL-cut) *na* (with) *kisu* (knife) (3.8)

"Juma cut himself with a knife"

bunge (SG-parliament) **i**-*li-ji-patia* (3SG-PAST-REFL-award) *nyongeza* (increment) *kubwa* (huge) *ya* (of) *mshahara* (salary)

"Parliament awarded itself a huge salary increment"

d) Locational suffix: Swahili rarely employs affixal case with the exception of the suffix *-ni* which forms a locational oblique when attached to a noun as example 3.9 shows:

Wa-fungwa	**wa**-*na-enda*	*mahakama-ni*	(3.9)
PL-convict	3PL-PRES-go	court-LOC	

"The convicts are going to court"

e) Count/Mass distinction: This provides a good indicator for most abstract nouns e.g. *'uhuru' freedom*, and mass nouns such as *'maji' water*, *'dhahabu' gold* etc. This information is obtained from the noun class prefix or agreement concords in verbs, adjectives, pronouns etc. In the first example 3.10, the noun class prefix *wa-* indicates plural, while in example 3.11, the noun class prefix indicates singular but the agreement concord in the pronoun shows that the noun is in plural form. Using the agreement concords for this purpose is especially useful in the case of nouns whose class prefix is always plural or singular, but that can take both forms.

Wa-toto	**wa**-*na-lala*	(3.10)
PL-child	3PL-PRES-sleep	

"The children are sleeping"

Wa-li-uza	0-*gari*	**z**-*ao*	(3.11)
3PL-PAST-sell	SG-car	PL-POSSPRON	

"They sold their cars"

f) Derivational affixes: A strong indicator for abstract nouns is their verb part or their attributive/adjective part. Many abstract nouns are derived from verbs and adjectives and thus derivational affixes offer vital clues. The nouns in example 3.12 are derived from adjectives *'huru' free* and *'zuri' good* to yield attributive nouns, while those in example 3.13 are derived from verbs *'tembea' walk* and

'kutana' meet yielding an action/activity and event noun, respectively.

U-huru	U-zuri	(3.12)
DER-free	DER-good	
freedom	goodness	
Ma-tembe-zi	*M-kutan-o*	(3.13)
PL-walk-DER	DER-meet-DER	
visits/travels	meeting	

2. Part of speech[22]

a) Preposition: Different types of prepositions typically co-occur with different types of nouns providing another important linguistic clue on semantic types of nouns. For example prepositions such as '*tangu' from*, '*hadi' till*, '*kabla-ya' before*, '*baada-ya' after*, '*mpaka' till/until* etc., take nouns denoting time (3.14), while '*ndani-ya' in/inside*, '*karibu-na' near*, '*kando-ya' beside/along*, '*katikati-ya' among/middle of* etc., occur with location types, though with some exceptions (3.15).

Maria	**a**-*ta-kaa*	*hoteli-ni*	*tangu*	*leo*	*hadi*	*kesho*	(3.14)
Maria	3SG-FUT-stay	hotel-LOC	from	today	till	tomorrow	

"Maria will stay in the hotel from today till tomorrow"

Paka	**a**-*me-lala*	*chini*	*ya*	*kiti*	(3.15)
SG-cat	3SG-PRES-sleep	under	of	chair	

"The cat is sleeping under the chair"

b) Numerals: These normally occur together with quantities or units of measure as example 3.16 shows. They also provide supplemental information useful for making the count/mass distinction, in the absence of concordial prefixes, as

[22] Here, focus is only on those parts-of-speech that may indicate the semantic type of nouns, rather than the enumeration of all parts of speech.

example 3.17 illustrates.

Juma	**a**-*li*-*nunua*	*lita*	*mbili*	*za*	*maziwa*	(3.16)
Juma	3SG-PAST-buy	litre	NUM-two	of	milk	

"Juma bought two litres of milk"

Fundi	**a**-*li*-*shona*	*nguo*	*tatu*	*jana*	(3.17)
SG-tailor	3SG-PAST-sew	SG-dress	NUM-three	yesterday	

"The tailor sewed three dresses yesterday"

3. Predicate-Argument selectional preferences: Grammatical relations provide a link between syntax and semantics. The verb and its direct dependents are central to the meaning of a sentence. By exploiting the grammatical relations between the verb and dependent nouns, i.e. subject and object, it is possible to gather semantic type information for the dependent nouns, taking into consideration the semantics of the verb. Nouns that are subjects or objects of the same verb (type) usually have some semantic similarities which may be generalized into a semantic type. In this study, given that the focus for WSD is on nouns, semantic properties (types) for Swahili verbs have been acquired via translation into English from two sources namely, WordNet and Levin's (1993) verb classes. In this way, the semantic type of the verb '*imba*' is determined to be **communication** by obtaining the tag associated with its English translation *sing*, from WordNet, and is a **sing-verb** using Levin's classes. From example 3.18 below, it is possible to infer that the noun '*msichana*' belongs to a semantic class of nouns that can communicate, since it is the subject of the communication verb '*imba*'. Likewise, the object '*wimbo*', the product of a communication process is an abstract noun as are all speech products.

Msichana	a-*li*-**imba**	**wimbo**	m-zuri	(3.18)
SG-girl	3SG-PAST-sing	PL-song	SG-ADJ-good	

"The girl sang a good song"

In many cases, the surface subject and direct object of a verb correspond to the first and second argument of the verb's semantic predicate. If they do not, e.g. in a passive sentence or due to the numerous verbal extensions applicable to the verb, the deep grammatical

Feature	**Values**	**#**
Noun prefix	$1-11(1/2, 3/4, 5a/6, 5/6, 6, 7/8, 9/10, 9/6, 11/10, 11/6, 11)$	1
Subject prefix	$1-6\ (1/2, 3/4, 5/6, 7/8, 9/10, 11)$	1
Location	0 (Loc suffix absent), 1 (Loc suffix present)	1
Reflexive	0 (Refl infix absent), 1 (Refl infix present)	1
Preposition	0 (absent PP), 1 (time PP), 2 (Loc PP), 3 (other PP)	1
Number	0 (Num absent), 1 (Num present)	1
Count/mass	1 (SG or PL only), 2 (both SG and PL)	1
Derived	0 (not derived), 1 (derived)	1
Pita	0 (Verb *pita* absent), 1 (Verb *pita* present)	1
WordNet classes	0 (Verb class absent), 1 (SUBJ verb class), 2 (OBJ verb class)	15
Levin classes	0 (Verb class absent), 1 (SUBJ verb class), 2 (OBJ verb class)	183

Table 3.3: Context features

relations determine the argument positions. By exploiting the SVO word order of Swahili, the arguments of the verbal predicate were obtained[23] from the analysis of individual words for those cases where the relevant syntactic tags were not generated directly by SWACGP.

Table 3.3 gives a summary of the features used in the study and their range of values.

3.3.2 Using the Self-Organizing Map to determine Semantic classes for WSD

The SOM is an unsupervised neural network method which maps complex and high-dimensional data onto a regular low-dimensional (two-dimensional) grid in an ordered fashion such that similar data inputs are, in general, located near each other (Kohonen 1995, Honkela 1997). This low-dimensional grid can then be effectively utilized to visualize and explore properties of the data.

[23] A Perl module was written that determines the subject and object of a verb using very simple syntax rules devised for this study. The verbal constructs covered include passive, stative, applicative, causative and their various combinations. This was deemed necessary since the Swahili verb rarely occurs in its most simple form, and ignoring the complex (those with verbal extensions + passive) form would have significantly reduced the cases from which this feature vector could be populated.

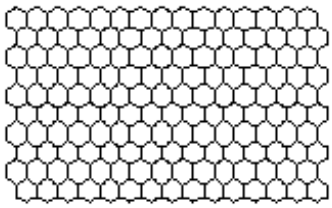

Figure 3.4: A hexagonal SOM grid

A SOM consists of neurons organized on a regular grid as shown in figure 3.4. Each neuron is a d-dimensional weight vector (codebook vector), where d is equal to the dimension of the input vectors. Each neuron serves as a model or prototype of a class of similar inputs. The neurons are connected to adjacent neurons by a neighbourhood function which determines the topology of the map i.e. the lattice structure (hexagonal or rectangular) and global map shape (sheet, toroid or cylinder). A unique property of the SOM is that it simultaneously forms a grouping (clustering) of the input data and performs a non-linear projection of the data set. This makes it an excellent tool for data mining due to the good visualization of any emergent categories obtained from the data. The SOM has been successfully used in a wide range of applications and domains. Examples include image processing, speech recognition, process control, economical analysis and industrial and medical diagnostics, amongst others. The SOM has also been used extensively in various NLP applications such as WEBSOM (Honkela et al. 1997).

For this study, the SOM is used solely as an exploratory tool to derive a semantic landscape of Swahili nouns, without focussing on its statistical or mathematical foundations. In this regard, the reader is referred to SOM literature such as Kohonen (1995) and Honkela (1997) for a comprehensive coverage of the SOM algorithm.

Creating the SOM

The 500 most frequent unambiguous nouns in the corpus were selected for this study. The criteria applied in selection was i) the noun must be listed in the TUKI dictionary, to ensure that its WordNet tag can be obtained via its translation, and ii) the selected nouns

must represent all of WordNet's 26 noun classes, so as to establish their correspondence to the noun clusters obtained by clustering the SOM. For each noun, all its occurrences were extracted from the corpus and analysed using SWATWOL which performs the initial pre-processing of the raw texts, morphological analysis as well as morphological disambiguation. Any remaining morphological ambiguities were left unresolved with the first given analysis assumed to be the correct one. To describe a noun, the contextual features given in table 3.3 were collected from a 5-word context window, two words on either side of the noun. Occurrences of nouns in idiomatic expressions as tagged by SWATWOL were ignored, and feature extraction was done within sentence boundaries i.e. features are collected only from the sentence in which the noun occurs. The resulting training data matrix, D, is formally described as follows:

Let N be the set of nouns, $n_i \in N, i = 1 \ldots 500$;

Let F be the set of context features, $f_j \in F, j = 1 \ldots 207$;

If $D = \{d_{i,j}\}$ represents the data vectors, then the value $d_{i,j}$ represents the frequency of feature f_j within the context of noun n_i and is a measure of how typical the jth feature is within the context of the particular noun, n_i. All the data vectors have been normalized by the total occurrences for each word.

Obtaining semantic classes by clustering the SOM

The SOM toolbox was used to organize the data vectors D and visualize the word categories. The organisation of the data is depicted in the distance matrix shown in figure 3.5. High values on the distance matrix (black color) denote large distances between neighbouring units, and represent cluster boundaries while the light areas on the map correspond to clusters. The largest cluster appears in the middle section of the lattice. Other smaller clusters are scattered around it and others are found on the right side of the lattice.

WordNet	Label	WordNet	Label
Time	a	Location	n
Substance	b	Group	o
State	c	Food	p
Shape	d	Feeling	q
Relation	e	Event	r
Quantity	f	Communication	s
Process	g	Cognition	t
Possession	h	Body	u
Plant	i	Attribute	v
Phenomenon	j	Artifact	w
Person	k	Animal	x
Object	l	Act	y
Motive	m	Tops	z

Table 3.4: WordNet labels

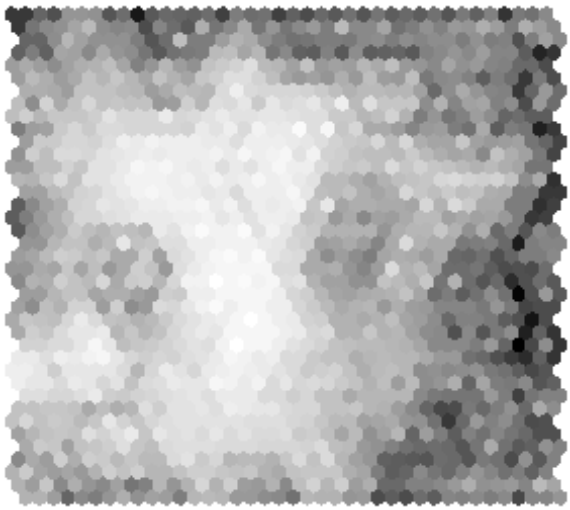

Figure 3.5: Distance matrix

For each noun, the best-matching map unit (bmu) was obtained by locating the model vector that most closely resembles that of the data. The word label (or its corresponding WordNet tag) was then written onto the map unit corresponding to the bmu, as shown in figures 3.6 and 3.7. For the latter, WordNet classes have been labelled alphabetically for visibility reasons, and table 3.4 shows each tag with its label (A-Z).

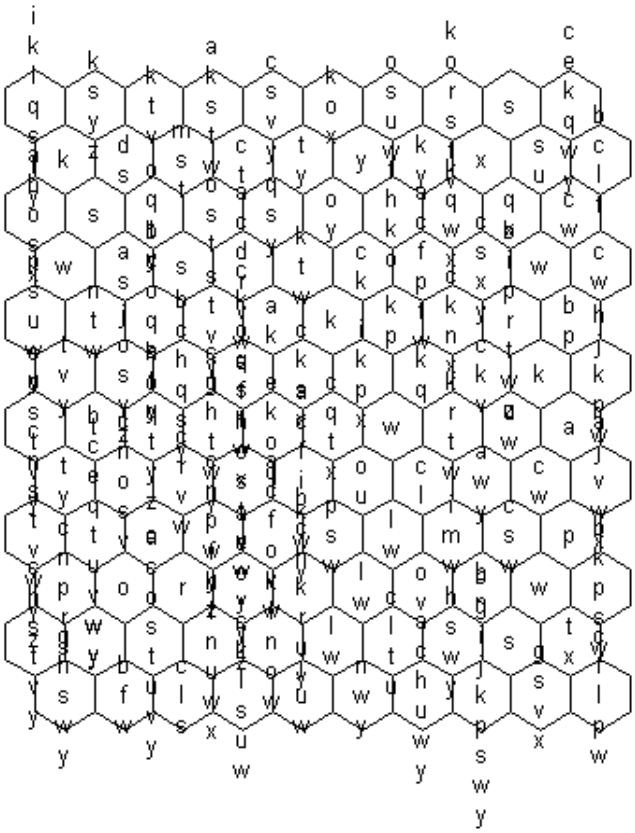

Figure 3.6: Map labelled with WordNet tags

To get a definite clustering of the data, MBC was used to cluster the SOM codebook vectors[24] (Banfield & Raftery 1993, Fraley & Raftery 2002). As one aim of using the SOM algorithm was to determine the correspondence between WordNet noun classes and induced categories, a clustering technique that automatically determines the optimal (best) number of clusters for the given data was preferred over one where this number is required as an argument, such as the k-means clustering algorithm. This way, the 'true' number of clusters inducible using Swahili features would be determined from the data itself rather than have this number chosen subjectively. The MBC algorithm requires as one of its arguments, the maximum number of clusters it should find. This was specified as 26 since

[24] Rather than cluster the data directly, the SOM has been used as an intermediate phase to reduce the computational complexity of clustering. Vesanto & Alhoniemi (2000) validate using such a two-level approach by showing that the two methods achieve comparable clustering results.

by assuming a one-to-one correspondence between WordNet and the learned categories, the data would organise into 26 clusters. If less than 26 clusters were obtained, then it would be possible to determine which WordNet classes have been split up and which ones combined to form new classes on the basis of Swahili linguistic evidence. This is done by analysing the properties of the new clusters. The clustering results are shown in figure 3.7. By comparing figures 3.6 and 3.7, and taking into account the component maps[25] (figure 3.8), individual units in the clusters were analysed in depth to identify their member nouns and semantic properties in order to determine which semantic classes can be deduced directly from the data. Table 3.5 shows example words derived from map units from selected clusters as indicated in figure 3.7. Abstract nouns appear largely on the left half of the lattice (clusters A, B, F) while concrete nouns are found mainly on the right half and bottom parts of the lattice (humans (C), food/substances (D), artifacts/dress (E) and locations (G). Some WordNet classes were consequently reclassified as shown in table 3.6.

[25] Visualizations of the component planes show what values the prototype vectors of the map units have for different vector components (features).

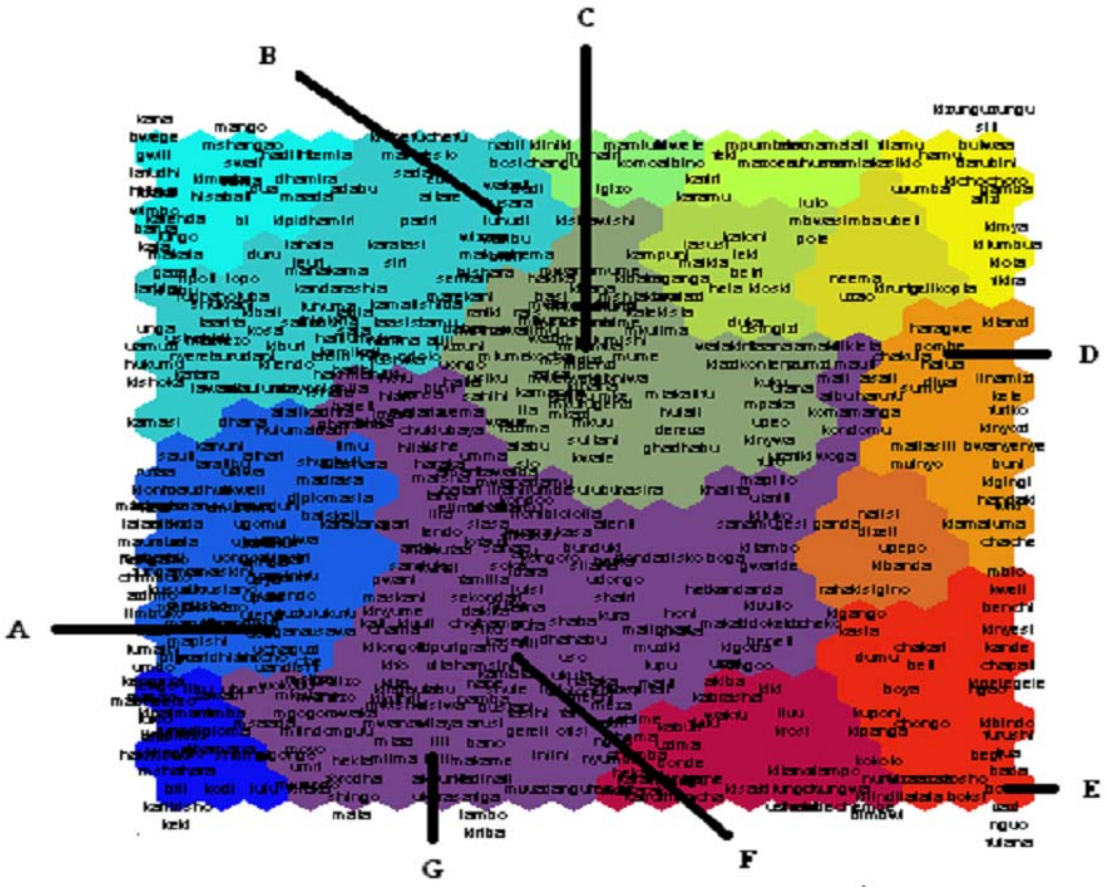

Figure 3.7: Map clustered into 15 classes using model-based clustering. Swahili words are used to label map units which correspond to their BMU.

The final set of semantic classes obtained via SOM clustering is shown in table 3.7. These classes form the basis of the WSD system as they are mutually exclusive and distinguishable using linguistic evidence derived from Swahili data as spelt out in section 3.3.

Data Acquisition and Annotation

Annotated training data for each of the identified semantic classes has to be obtained. The data is used to learn Bayesian classifiers for WSD using the BMT in a supervised learning setting. Each training example must therefore be tagged with the label of its corresponding class. During cluster analysis, member nouns for each cluster were identified,

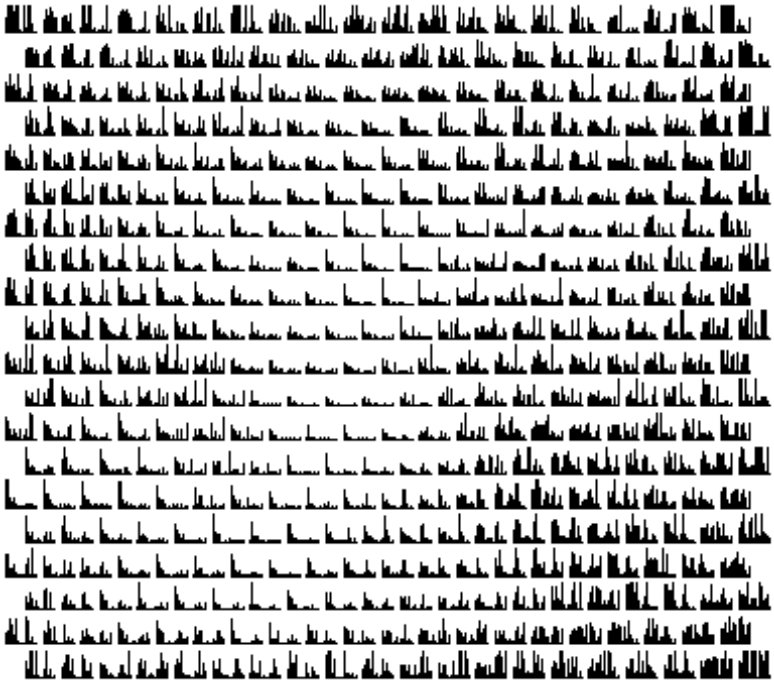

Figure 3.8: Bar chart visualization of the prototype vectors for individual map units.

Cluster	Swahili words and their translations
A	*uhai, utajiri, upendo, ujuzi, upana, uchaguzi, uandishi, wokovu* life, wealth, love, expertise, width, selection, authorship, salvation
B	*wakati, busara, juhudi, wajibu, bidii* time, good judgement, effort, responsibility, effort
C	*msaidizi, mwenyekiti, katibu, mkazi* assistant, chairperson, secretary, inhabitant
D	*haragwe, pombe, halua, divai* bean, beer, sweetmeat, wine
E	*furushi, jua, bohari, vazi, nguo, fulana, kizaazaa, kibindo* bundle of clothes, sun, warehouse, clothing, cloth, undershirt, chaos, loin cloth pocket
F	*dakika, siku, gramu, hamsini, nane, namba* minute, day, gram, fifty, eight, number
G	*kisiwa, wilaya, jiji, kijiji, kitongoji, ofisi* island, district, city, village, small village (hamlet), office

Table 3.5: Examples of words taken from different clusters

WordNet Tag	**Re-classified Tag**	**Example Words**
Object	Location	Mountain, River, Lake
Possession	Location	Land
Possession	Artifact	Painting
Possession	Substance	Gold, Silver
Artifact	Location	House, Pool, Shop
Communication	Artifact	Book, Magazine, Newspaper
Communication	Abstract	Song, Insult, Prayer
Cognition, Event, Feeling, Motive, Process	Abstract	Sight, Marriage, Fear
Relation, Shape, State, Act, Attribute	Abstract	Peace, Beauty, Dance

Table 3.6: Re-classified WordNet tags

Class	**WordNet correspondence**
Human	Person, Tops
Animal	Animal
Location	Location, Object, Artifact, Possession
Time	Time
Unit	Quantity
Substance	Substance
Body	Body
Food	Food
Plant	Plant
Abstract	State, Shape, Relation, Process, Phenomenon, Motive, Feeling, Event, Communication, Cognition, Attribute, Act
Artifact	Possession, Communication, Artifact
Money	Possession
Dress	Artifact
Vehicle	Artifact
Container	Artifact

Table 3.7: Semantic classes derived from Swahili data

Class	Nouns	Occurrences
Abstract	93	70553
Animal	37	1963
Artifact	49	10238
Container	12	605
Dress	13	1622
Food	42	3718
Human	32	45993
Institution	32	31587
Location	46	23758
Money	20	6173
Plant	34	334
Substance	38	3923
Time	30	25349
Unit	29	1580
Vehicle	12	2906

Table 3.8: Automatic annotation of data: Unambiguous nouns and their corpus occurrences

and these are used to supply the training examples. All occurrences for each member noun of a given class were extracted from the corpus, processed individually using SALAMA, relevant features extracted and coded[26] in a format suitable for BMT. The class label was added to each example. All the labelled examples for all nouns of a given class were combined into one training file for that class i.e. 15 different training files, one for each of the 15 classes are created. Table 3.8 shows, for each class, the number of unambiguous nouns occurring in the corpus as well as their combined corpus occurrences, from which labelled data is obtained. Due to the time expense of obtaining a gold standard for testing, test data for each class was obtained by deleting the class label for 10% of the training data and reserving these examples as test data. Nonetheless, a small gold standard was prepared for key ambiguity types by hand-tagging $2,528$ occurrences of sixteen ambiguous words and using these to test the performance of the WSD algorithm on actual ambiguous words.

[26] The training data for the SOM training consisted of a normalized vector for each training word. For the Bayesian learning, each training example represents an individual occurence of a training word in the corpus.

3.4 Analysis of Lexical Translational Ambiguity in Swahili Nouns

With a compact set of semantic categories for meaning representation identified, the ambiguity types prevalent in the language pair are determined. This is done by analysing the ambiguous nouns to reveal the semantic types associated with each of their readings. The ambiguous nouns are then sorted into ambiguity groups, where an ambiguity group comprises nouns that share the same type of ambiguity and therefore rely on similar semantic distinction criteria for disambiguation. These groups form the basis for an 'ambiguity-type driven' approach to WSD where Bayesian classifiers are constructed for each of the groups as discussed in section 3.5. The following subsections discuss the ambiguity analysis process in further detail.

3.4.1 Ambiguity Prevalence

Many words are semantically ambiguous, referring to more than one concept. In addition, words can be ambiguous in different ways. Some words are ambiguous between highly related senses whose semantic relationship is systematic. For example, the *stroke* meaning of the word '*bakora*' is derived from the second meaning *cane* or *walking stick* and refers to the action of using a cane. On the other hand, the two meanings of a word like '*mkesha*', *eve/vigil* and *sparrow* are semantically unrelated, and seem to share the same written form purely by chance. The linguistic literature makes a distinction between these two types of ambiguity, with the former referred to as polysemy and the latter homonymy (Lyons 1977, Cruse 1986). Most standard dictionaries reflect this distinction between word meanings and word senses, where word meanings correspond to different lexical entries, and related word senses are contained within a single entry. TUKI's Swahili-English bilingual dictionary adheres to this format where ambiguous words are listed as separate entries if considered homonyms or as single entries with numbered senses, in the case of polysemes.

In this study, a noun is determined to be ambiguous if marked either as a homonym or polyseme in the TUKI dictionary, using the above format. Ambiguous nouns make up 21%

Part of Speech	Ambiguity	% of Homonyms	% of Polysemes
Nouns	21%	40%	60%
Verbs	31%	19%	81%
Adjectives	14%	4%	96%
Adverbs	10%	10%	90%

Table 3.9: Translational ambiguity prevalence in Swahili

n-way ambiguity	% (nouns)
2-way	72
3-way	20
4-way	6
$\geq$5-way	2

Table 3.10: Noun ambiguity

of all listed nouns, as shown in table 3.9, with 60% of these being polysemous. While this distinction is not critical for WSD as it is for applications such as IE or IR, it nonetheless provides important information regarding sense granularity. Whether the individual senses of an ambiguous word are coarse or fine-grained determines the level of difficulty of the disambiguation task and this directly or indirectly influences various aspects of algorithm design such as the types of relevant disambiguation information to use, whether to adopt a general disambiguation algorithm for certain words or to build individual word-specific disambiguators or even how to evaluate WSD performance (Resnik & Yarowsky 1999).

3.4.2 Ambiguity Types

Ambiguity types important for WSD were identified by processing all ambiguous nouns listed as such in the TUKI dictionary as follows:

1. Select all nouns that are two-way[27] ambiguous. These form the majority as shown in table 3.10.

[27] Two-way ambiguity is taken as the base case, and all other n-way ambiguities are combinations of the ambiguity types obtained from analyzing two-way ambiguity.

Tag I	**Tag II**	**Word**	**Reading I**	**Reading II**
Human	Artifact	*Gumegume*	Worthless person	Flint gun
		Kiongozi	Leader, guide	Manual, handbook
		Mlezi	Guardian, custodian	Cot
	Location	*Pepo*	Demon, spirit	Paradise
		Bucha	Butcher	Butchery
	Animal	*Kirukanjia*	Prostitute	Nightjar
		Sungusungu	Homeguard, vigilante	Black ant
		Mkunga	Midwife	Eel
	Food	*Jini*	Genie, wicked person	Gin
		Nyanya	Grandmother	Tomato
	Plant	*Mtini*	Clown, buffoon	Fig tree
	Time	*Juma*	Name of person	Week
	Abstract	*Kichaa*	Lunatic	Lunacy
		Mwanga	Wizard	Light
		Nyange	Fool, moron	Noise

Table 3.11: Ambiguity Group: Human

2. For each noun, obtain the WordNet tag[28] corresponding to each of its English translations (meanings /senses). The WordNet tags are used to represent the initial meanings of the English readings. The second stage involves retagging the nouns using the new class labels where applicable.

3. Split the nouns into two groups depending on whether their English translations have different tags or not. Example words are shown in tables 3.11-3.17 and 3.18 respectively. The rows in the first two columns of each individual table represent the types of ambiguities that the Bayesian classifiers have to learn to disambiguate. This division (of nouns into two groups) clearly illustrates what can be accomplished by the means available from raw textual data, with respect to WSD. In this case, the WSD solution covers only those cases where the English translations have different tags. Construction of Bayesian classifiers based on these classes is discussed in detail in section 3.5.

[28] Where the English reading is ambiguous, the MFS (listed first in the noun's entry) is chosen as the correct translation of the Swahili word. There are however, cases where it goes wrong.

Tag I	**Tag II**	**Word**	**Reading I**	**Reading II**
Animal	Vehicle	*Ndege*	Bird	Aeroplane
	Artifact	*Simu*	Sardine, sprat	Telephone
	Body	*Koo*	Hen, breeding animal	Throat, gullet
	Food	*Tembe*	Hen	Tablet
		Kima	Black monkey	Minced meat
	Abstract	*Swala*	Gazelle	Prayer
		Goma	Hard-skinned fish	Stick dance
		Kima	Black monkey	Price, value, rate
	Money	*Mbango*	Warthog	Money
	Dress	*Buibui*	Spider	Purdah, veil
	Time	*Mkesha*	Sparrow	Eve, vigil
	Container	*Chungu*	Black ants	Pot
	Location	*Korongo*	Stork, crane	Gulley, ravine
		Barabara	Crowned hornbill	Highway, road, street
		Paa	Gazelle	Roof

Table 3.12: Ambiguity Group: Animal

Tag I	**Tag II**	**Word**	**Reading I**	**Reading II**
Location	Artifact	*Mkoa*	Province, region	Metal bar
		Komeo	Creek, inlet	Bolt, latch
		Mto	River	Pillow
	Food	*Kiwanda*	Factory	Omelette
		Tembe	House	Tablet
	Time	*Mwezi*	Moon	Month
		Magharibi	West	Sunset
	Body	*Ziwa*	Lake	Breast
	Plant	*Kambi*	Camp	Cambium
		Ua	Yard	Flower
	Vehicle	*Dau*	Pool	Dhow, sailboat
	Money	*Pango*	Cave	Rent
	Abstract	*Njia*	Road	Method, means
		Kitende	Residence, abode	Elephantiasis

Table 3.13: Ambiguity Group: Location

Tag I	Tag II	Word	Reading I	Reading II
Abstract	Plant	*Chacha*	Ballroom dance	Grass
		Dege	Convulsions, stomach pain	Fern
		Mti	Scrofula, gangrene	Tree
	Artifact	*Usukani*	Leadership	Rudder, steering wheel
		Breki	Break, recess	Brake
		Useja	Celibacy, bachelorhood	Collar
		Kifungo	Detention	Button
		Mwiko	Taboo, totem	Wooden spoon
	Time	*Magharibi*	(sunset) Prayer	Sunset
		Alasiri	(afternoon) Prayer	Afternoon
		Alfajiri	(morning) Prayer	Morning
	Food	*Zambarau*	Purple	Damson plum
		Bia	Cooperation, agreement	Beer
	Substance	*Madadi*	Assistance, support	Opium
		Ambo	Disease	Gum, glue
	Dress	*Doria*	Security patrol	Organdie, muslin
		Dibaji	Preface, preamble	Woollen/silk material
	Container	*Tusi*	Insult, abusive remark	Coffin, bier
	Body	*Sini*	Complexion, shape	Gum (of teeth)
	Vehicle	*Jipu*	Boil, abscess	Jeep

Table 3.14: Ambiguity Group: Abstract

Tag I	Tag II	Word	Reading I	Reading II
Artifact	Body	*Chupa*	Bottle	Amniotic membrane
		Kiko	Tobacco pipe, briar	Elbow
		Sini	Porcelain, chinaware	Gum (of teeth)
	Food	*Sindano*	Needle	Long thin rice
		Kiwanda	Weaving slivers	Omelette
		Pau	Rafter	Bread
	Container	*Kadi*	Card	Caddy
		Waya	Wire	Baking dish
	Unit	*Chembe*	Spear head	Iota, morsel
	Substance	*Saruji*	Saddle	Cement, concrete
	Time	*Saa*	Clock, watch	Hour
	Money	*Bakora*	Walking stick, malacca cane	Apprenticeship fees

Table 3.15: Ambiguity Group: Artifact

Tag I	Tag II	Word	Reading I	Reading II
Institution	Money	*Dola*	State, government	Dollar, buck
	Substance	*Ukoo*	Clan, kinship, family	Filth, dirt
	Body	*Bodi*	Board	Body

Table 3.16: Ambiguity Group: Institution

Tag I	**Tag II**	**Word**	**Reading I**	**Reading II**
Time	Plant	*Chaka*	Hot season	Thicket
Substance	Plant	*Tete*	Slag, dross	Reed
Body	Unit	*Futi*	Knee	Foot
Dress	Money	*Kilemba*	Turban	Dowry/gratuity/bribe

Table 3.17: Ambiguity Group: Time, Plant, Substance, Body, Unit, Dress, Money

Tag I	**Tag II**	**Word**	**Reading I**	**Reading II**
Human	Human	*Wakili*	Advocate, counsel	Commissioner
		Mshenga	Agent, go-between	Intermediary
Animal	Animal	*Nyoka*	Snake	Worm
		Mamba	Crocodile, alligator	Black mamba
Location	Location	*Kasri*	Mansion	Palace
		Jangwa	Desert	Wilderness
Abstract	Abstract	*Kofi*	Dance	Slap
		Radhi	Contentment, satisfaction	Apology, pardon
Time	Time	*Mchana*	Daytime	Afternoon
		Juzi	Day (before yesterday)	Day (few days ago)
Artifact	Artifact	*Fimbo*	Stick, mace	Walking stick
		Upanga	Sword	Long wooden knife
Food	Food	*Mkate*	Bread	Tobacco cake
Body	Body	*Ondo*	Knee	Leg, foot
Substance	Substance	*Kifusi*	Rubble	Debris
Container	Container	*Jeneza*	Bier	Coffin
Dress	Dress	*Kanzu*	Cassock	Gown

Table 3.18: Nouns with similar noun tag for both readings

As shown in table 3.18, most cases where the English translations have a similar tag reflect very highly related meanings, with some readings being specializations of the other reading. For example, WordNet defines a *palace* as a type of *mansion* while *rubble* and *debris* are near synonyms. Most of these cases represent very fine-grained sense distinctions that cannot be handled by the broad semantic classes identified for Swahili, and have therefore not been addressed by the WSD solution.

3.5 Bayesian Classifiers for WSD

In section 3.3.2, semantic classes representing the most important semantic distinctions for WSD within a Swahili-English MT context were identified. In this setup, WSD has been recast as a classification problem and disambiguation consequently involves determining semantic class membership between two or more competing classes, where each class represents a different sense of the ambiguous word. The English reading associated with the winning class, as determined via WordNet association, is then chosen as the disambiguated sense of the ambiguous Swahili noun, thereby achieving WSD (TWS).

Michie et al. (1994) define the task of classification as any context in which a decision or prediction is made based on currently known information, using some classification procedure. The construction of the classification procedure is one of the most common learning tasks which has been variously addressed using statistical, ML and neural network approaches. In this study, ML has been used to induce the WSD classification procedure. ML has been defined by numerous authors: Weiss & Kulikowski (1991) refer to a learning system simply as a computer program that makes decisions based on the accumulated experience contained in successfully solved cases, while Mitchell (1997) gives a more formal definition: "A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E". The common thread in these two definitions is that the computer system *learns* how to *make decisions on a new instance* of a certain *task* based on

accumulated *experience derived from solved cases.* The system uses a given *performance evaluation measure* to *improve* its performance, with increased experience.

The fundamental goal therefore of empirical learning is to extract a decision rule i.e. learn a target function, from sample data, that will be applicable to new instances of data. To do this, a suitable representation of the target function has to be selected, and usually this is a general model such as a neural net, a discriminant function, a decision tree, a probabilistic model etc. An algorithm that is applicable to the chosen representation model is then used to learn the target function from the data samples. Learning in this sense entails selecting the model parameters and adapting them accordingly to obtain a generalized function that not only fits the sample data well, but also makes correct predictions on new samples (Weiss & Kulikowski 1991). Figure 3.9 clearly summarizes the process of learning a classification procedure (classifier) and using it to predict the class of a new instance.

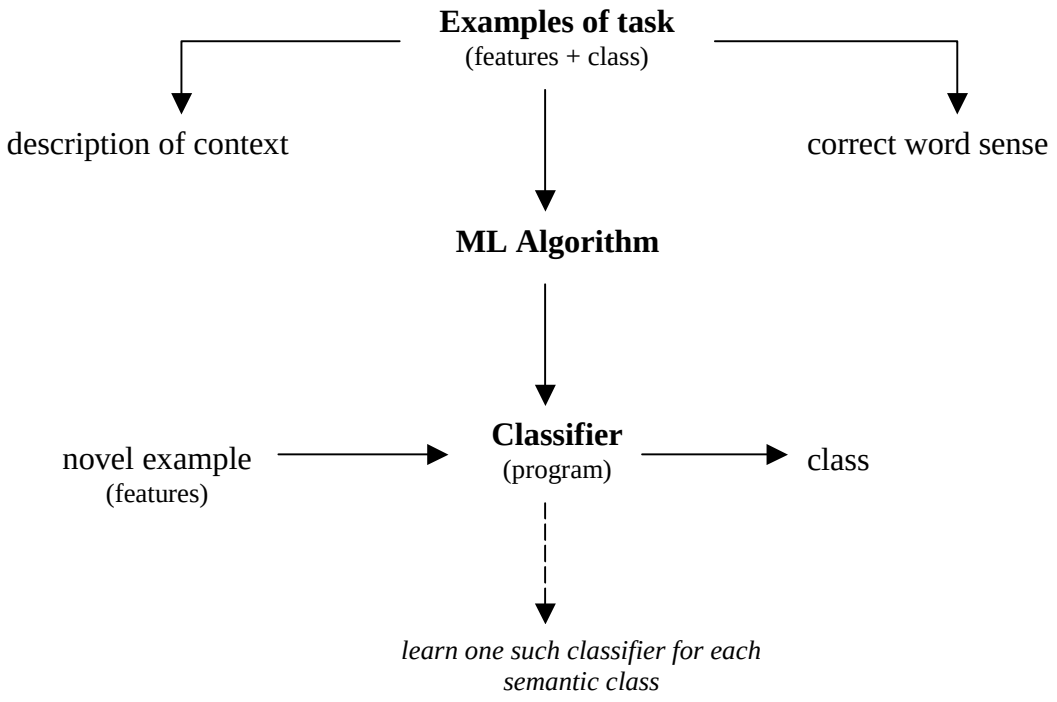

Figure 3.9: Learning a classification system

In this study, probabilistic models have been chosen to represent the target function (knowledge to be learned). A brief introduction on the basics of probabilistic models is presented in section 3.5.1, followed by a discussion on how to use such models as classifiers to perform WSD.

3.5.1 Probabilistic Models

Statistics provides a way to make inferences about a population from just a sample of that population, rather than having to study the entire population. This is achieved by acquiring a random sample of the population and identifying observable attributes or features of interest in this population. Random variables are used to represent these features. Next, the event space is identified. The event space is the total collection of all the events associated with this sample, where an event refers to any possible outcome of an experiment, or state of a process at a given observation time. A particular instantiation of values for the set of random variables therefore describes a particular event in the sample space. The numerical characteristics of the population under study can then be known via statistical inference, from the parameters that describe each event in the event space.

The dependencies existing among the random variables together with the estimated parameter values associated with each distinct event in the event space are represented in a probabilistic model. The probability distribution over this joint event space is called the *joint probability distribution* and specifies the probability of occurrence for any distinct event. For example, if X is an arbitrary set of random variables $x_1 \cdots x_n$, and each variable x_i can assume any value in the set $V(x_i)$, the event space of the set of variables X is defined as the cross-product $V(x_1) \times V(x_2) \cdots \times V(x_n)$. The probability that a specific event i.e. variable bindings for the tuple $< x_1 \cdots x_n >$ will occur, can then be determined from the joint probability distribution. A probabilistic model thus consists of a parametric form (that describes the dependencies among the features) and parameter estimates (that tell how likely each possible event is to occur) and such a model can be used as a classifier to identify the most probable sense of an ambiguous word given the context in which it appears.

In the context of this study, the task is to learn a probabilistic model for WSD. The training data defines the sample space. The events are Swahili sentences that contain unambiguous words that are representative of the semantic classes whose properties are to be learned. The contextual features shown in table 3.3 are the random variables that describe the events (sentences). The parameters of the model describe how likely it is to observe a particular feature vector i.e. instantiation of the feature variables, for any given sentence. The learning problem thus involves determining the parametric form of the probabilistic model and obtaining estimates for the parameters from the training data. This yields a fully defined joint probability distribution. This is discussed in section 3.5.2.

3.5.2 Bayesian Learning

The goal of ML in general is to determine the best (most probable) hypothesis (target function), referred to as the Maximum a Posteriori (MAP) hypothesis h_{MAP}, out of a set of possible hypotheses H, while minimizing the overall error rate. h_{MAP} is selected as the hypothesis $h \in H$, that has the highest posterior probability, denoted as $P(h \mid D)$, as determined from some observed data (evidence) D, and any prior information about the probabilities of the hypotheses in H (Mitchell 1997). Hypothesis h_i is selected as the most probable hypothesis given the data, based on equation 3.19.

$$P(h_i \mid D) > P(h_j \mid D) \text{ for all } i \neq j \tag{3.19}$$

Computing the posterior probability of a hypothesis requires an enormous sample space from which fully-specified probability data for all the statistical dependencies among the feature variables can be determined. Since these probabilities are derived from limited training data that is not sufficiently exhaustive in terms of feature combinations, computation of the posterior probabilities of different hypotheses is a challenge. This problem is alleviated by Bayes theorem (equation 3.20) which relates the posterior probability of a hypothesis to the conditional probability of observed data for a specific hypothesis, denoted $P(D \mid h)$, and to the prior probability of the hypothesis, $P(h)$.

$$P(h \mid D) = \frac{P(D \mid h) \times P(h)}{P(D)} \tag{3.20}$$

Using Bayes theorem, the MAP hypothesis is selected as the one with the highest posterior probability as shown in equation 3.21

$$\begin{aligned} h_{MAP} &\equiv \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h) \times P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h) \times P(h) \end{aligned} \tag{3.21}$$

In the final step, the constant term $P(D)$ is dropped from the equation as $\arg\max h$ does not depend on it.

Therefore, determining the best hypothesis for the data, which is subsequently used as the WSD classifier, requires the estimation of $P(D \mid h)$ and the prior probabilities $P(h)$ for all the hypotheses in H. Estimating $P(D \mid h)$ is a non-trivial learning task and various methods and techniques exist for obtaining this estimate from the training data. Each of these methods makes different assumptions about the characteristics of the hypotheses and this dramatically reduces the amount of information necessary to specify the full joint probability distribution. This in turn simplifies the acquisition of these conditional probabilities. In this study, two Bayesian methods namely, the Naïve Bayes (NB) classifier and BBNs are used to learn probabilistic classifiers for Swahili WSD. They differ in the independence assumptions that they make as discussed in the following sections. The comparative performance of classifiers based on these two variations is presented in chapter 4.

Naïve Bayes

The NB learner assumes that all the contextual features are independent given the class variable, and as such the parametric form is always the same as depicted in figure 3.10. The learning task therefore involves obtaining the parameter estimates from the data, with no explicit search for the best hypothesis (Mitchell 1997).

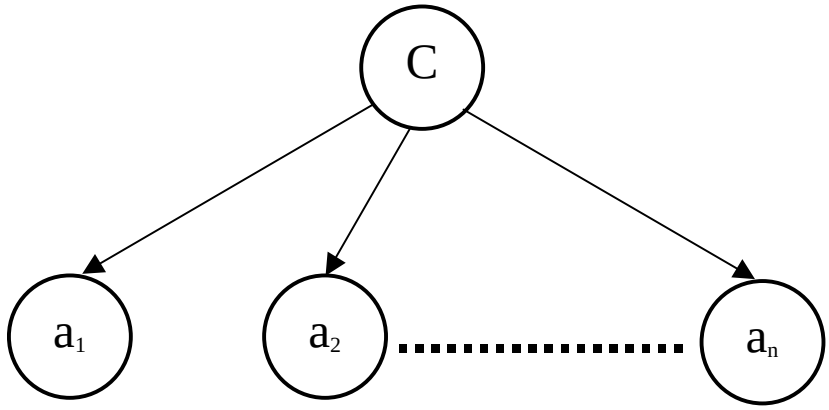

Figure 3.10: Naïve Bayes model showing absolute independence of feature variables $a_1 \cdots a_n$ given the class variable C.

The goal of the classifier is to assign the most probable class c_{MAP} out of a set of predefined classes $C = \{c_{1,}c_2, \cdots, c_k\}$, given a test instance (evidence) $\mathbf{e}$ according to equation 3.22.

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid e) \tag{3.22}$$

Since $\mathbf{e}$ is described by the feature vector $\{a_1, a_2, \cdots, a_n\}$, equation 3.22 can be rewritten as

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid a_1, a_2, \cdots, a_n) \tag{3.23}$$

using Bayes theorem, equation 3.23 is rewritten as

$$c_{MAP} = \arg\max_{c_i \in C} \frac{P(a_1, a_2, \cdots, a_n \mid c_i)P(c_i)}{P(a_1, a_2, \cdots, a_n)} \tag{3.24a}$$

$$= \arg\max_{c_i \in C} P(a_1, a_2, \cdots, a_n \mid c_i)P(c_i) \tag{3.24b}$$

The NB learning task thus requires the estimation of $P(a_1, a_2, \cdots, a_n \mid c_i)$, the conditional probabilities, and the priors for each of the classes in C, $P(c_i)$. In the absence of any additional knowledge, $P(c_i)$ can be computed as the proportion of each class in the training data. However, this will only be a valid estimate if the training data was ob-

tained via truly random sampling. Estimating the different $P(a_1, a_2, \cdots, a_n \mid c_i)$ requires enormous amounts of training data to ensure that each combination of the attribute values occurs a statistically sufficient number of times in order to obtain reliable estimates, a requirement which is practically not feasible for most applications. However, with the NB assumption that the attributes $a_1, \cdots, a_n$ are conditionally independent given the class variable, this conditional probability $P(a_1, a_2, \cdots, a_n \mid c_i)$, is simply computed as the product of the probabilities for individual attributes given the class. These probabilities are much easier to estimate as shown in equation 3.25:

$$P(a_1, a_2, \cdots, a_n \mid c_i) = \prod_{j=1}^{n} P(a_j \mid c_i) \tag{3.25}$$

substituting this term (equation 3.25) into 3.24b, then the NB classifier determines the most probable class given the test instance as that which maximizes equation 3.26.

$$c_{MAP} = \arg\max_{c_i \in C} \prod_{j=1}^{n} P(a_j \mid c_i) P(c_i) \tag{3.26}$$

Despite this simplification in its application of Bayes rule, the NB classifier has been used extensively in language learning applications with numerous researchers reporting that it performs just as well as other learning algorithms such as decision trees and artificial neural networks and even outperforms them in some cases (Michie et al. 1994).

Bayesian Belief Networks

The absolute conditional independence assumption made by the NB classifier rarely holds in practice, much less for natural language data where there are inherent dependencies among language units. This may sometimes result in degraded performance for some applications where this assumption clearly does not hold, such as in this study where numerous dependencies exist amongst the feature variables selected. For example, there is a strong correlation between noun and subject prefixes in Swahili: given noun prefix 1/2, the subject prefix is always 1/2. Another example is noun prefix 3/4 and its associated subject prefix 3/4. The subject prefix and the subject-verb features are also dependent on each other. For example the subject prefix 1/2 necessarily co-occurs with verbs that typi-

cally require animate subjects. Other examples include subject prefix vs. reflexive markers, preceding preposition vs. locational suffix, noun/subject prefix vs. count/mass feature etc. Rather than assume absolute independence for all the feature variables as NB does, BBNs[29] allow stating of conditional independence assumptions for subsets of variables, an approach that is not as constraining as the simple NB, and thus better suited to modelling real applications.

A BBN is a data structure that represents the dependencies among sets of variables along with the corresponding conditional probabilities, resulting in a concise specification of the full joint probability distribution governing the variables. It is represented as a directed graph that consists of nodes and directed arcs (links). The random variables make up the network nodes while the directed arcs between nodes depict dependencies. The arcs represent the assertion that the variable is conditionally independent of its non-descendants in the network, given its immediate predecessors in the network. If there is a directed arc from node X to Y, X is said to be a parent of Y and this means that X has a direct influence on Y. For each node X in the network, there is an associated Conditional Probability Table (CPT) that describes the probability distribution of that variable given its parents. The parameters of a Bayesian network model M thus consist of probabilities of the form $P(X_i = x_k \mid \prod_i = \pi_j)$ where $\prod$ denotes the parents of variable X_i and π denotes their value configuration.

The BBNs therefore provide a compact and complete specification of the domain where the probability of any event $P(x_1, \cdots, x_n)$, can be calculated from the network as a product of the relevant elements of the CPTs. By exploiting conditional independence, BBNs simplify the specification of the joint probability distribution by requiring specification of only the individual CPTs for each variable which results in a great reduction in the number of probabilities that have to be estimated. Therefore to use BBNs for probabilistic inference, the network topology (nodes and arcs) and the CPTs for each variable have to be specified. For simple domains i.e. those with few variables and whose exact dependencies are known, this can be done by a domain expert. For more complex domains with

[29] Also referred to as belief networks, probabilistic networks or causal networks.

several variables whose interdependencies may not be clearly known, ML algorithms can be employed not only to estimate the conditional probabilities, but also learn the network topology, as briefly described in the following section. The induced probabilistic model corresponds to the best hypothesis for the data, which is then used as a classifier for WSD.

As Heckerman (1996) notes, by encoding the dependencies among all the variables, BBNs are able to cope well with incomplete data, a feature that is very useful when the feature variables are highly anti-correlated. In such cases, when one of the inputs is not observed in the data, most models will produce an inaccurate prediction because they do not encode the correlation between the input variables, unlike the BBNs which do. In addition, by facilitating combination of domain knowledge with data, BBNs offer a natural way to make the best of any prior knowledge to complement the data. This is very important especially when training data is scarce or expensive to acquire, as is the case with obtaining annotated training data for WSD. Given this ability to cope well with problems of incomplete and sparse data, and considering too the general success of BL at the WSD task as reported by several researchers (Mooney 1996, Ng 1997, Leacock et al. 1998), the WSD classifiers have been modelled as BBNs in this study.

Learning BBNs from Data

Using Bayesian networks for prediction requires computing the average posterior probability for the data D, given all the possible network structures for the domain (X), if using the full Bayesian approach. To compute this average, the full posterior distribution for all the possible models (network structures) (M), would have to be determined using equation 3.27. This presents a computation bottleneck due to the huge number of possible models which is more than exponential in n where n is the number of network nodes (domain variables) Heckerman (1996). For example, Myllymäki et al. (2002) state that the hugely underestimated number of possible Bayesian network structures for 20 variables is $1.6 * 10^{57}$!

$$P(M \mid D) = \frac{P(D \mid M)\, P(M)}{P(D)} \tag{3.27}$$

Two statistical approaches are usually employed to address this problem - selective model averaging which involves choosing a small set of 'good' models from among all possible models and assuming that they represent the domain exhaustively, and model selection, which chooses only one 'good' model and assumes that it is the best or correct model for the domain, ignoring all other possible models. Different criteria are used to determine what constitutes a 'good' model, and these are extensively discussed in the literature on learning with Bayesian networks (Dawid 1984, Howard & Matheson 1984, Spiegelhalter et al. 1993). Despite this oversimplification of the full Bayesian approach, various researchers have shown experimentally that both model selection and model averaging often achieve accurate predictions (Cooper & Herskovits 1992, Aliferis & Cooper 1994, Heckerman, Mamdani & Wellman 1995, Madigan et al. 1996).

With the model selection criterion selected, the next task is that of using it to select a good model from all the possible models. Standard Bayesian selection takes the best model M', to be the one that is most probable for the data i.e. the model which yields the maximum posterior probability for the data, as shown in equation 3.28. Note that the constant term $P(D)$ in equation 3.27 has been ignored here since the $\arg\max$ does not depend on it.

$$M' = \arg\max_{M} P\left(M \mid D\right) = \arg\max_{M} P\left(D \mid M\right) P(M) \tag{3.28}$$

Finding the most probable model has been described as a NP-hard problem by Myllymäki et al. (2002) citing Chickering et al. (1994) and consequently, heuristic search algorithms are used in practice to find the most probable model for the data. Most search methods for Bayesian networks start with an initial network e.g. the empty network or a random graph, and make successive arc changes to this network retaining only those changes that yield a maximum positive increase in the probability of the model. Common search algorithms include greedy search, greedy search with restarts, best-first search and Monte-Carlo methods[30].

[30] For a more detailed discussion on the specifics of learning Bayesian networks from data, see Buntine (1991), Bernando & Smith (1994), Heckerman, Geiger & Chickering (1995), Jensen (1996), Heckerman (1996) and Pearl (2000).

The Bayesian Modelling Tools used for this study employ a combination of stochastic and greedy search heuristics to select the best model for the data, using the model selection criterion shown in equation 3.29.

$$P(D \mid M) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \tag{3.29}$$

where Γ denotes the gamma function, n is the number of variables in M, q_i is the number of value configurations for the parents of variable X_i, r_i is the number of values of X_i, $N_{ijk}, i = 1 \ldots n, j = 1 \ldots q_i, k = 1 \ldots r_i$ is the number of rows in D where variable X_i has value x_k and the parents $\prod_i$ of X_i have a value configuration π_j and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. The constants N'_{ijk} are the hyperparameters determining the prior distribution $P(M)$. A uniform prior distribution $P(M)$ over the models is assumed. Kontkanen et al. (2000) and Myllymäki et al. (2002) discuss the theoretical foundations and implementation specifics of the Bayesian Modelling Tools.

3.6 Supervised Learning of Bayesian Classifiers for WSD from annotated data

In the classification paradigm of supervised ML, a classification procedure is induced from a set of data for which the true classes are known, for a set of pre-defined classes $\{1, \cdots, K\}$. For WSD, learning such a classification procedure requires the availability of sense-tagged data, where each training example x_i is described by a feature vector and a corresponding class label. The feature vector comprises of attribute-value pairs, where the attributes are those contextual clues important for classification. The supervised learning task, as discussed in the preceding section, thus involves capturing important dependencies in the training data and representing these in a parametric model, from where the joint probability distribution can be defined. Once all the required model parameters have been estimated, the learned model can then be used as a classifier for WSD i.e. given a particular instantiation of the feature variables for a test sentence, the classifier predicts the value of the classification variable. It is the expectation that the learned classifier should perform

Set	Description	#
A	All Features	207
B	Morphological, POS, Co-occurrence, WordNet-based Selectional Preferences	24
C	WordNet-based Selectional Preferences	15
D	Levin-based Selectional Preferences	183
E	Morphological, POS, Co-occurrence, Levin-based Selectional Preferences	192
F	Morphological, POS, Co-occurrence	9

Table 3.19: Context feature sets

well in classifying test examples, and its prediction accuracy is used to measure how well it has been able to generalize from the training data to unseen data. The set of classifiers to be learned, together with their corresponding training data sets were defined in section 3.3.2. The Bayesian Modelling Tools are used to learn the Bayesian classifiers.

3.6.1 Training Parameters and Conditions

As Agirre & Martinez (2001) state, certain types of information are more effective than others in disambiguating certain types of ambiguities. In cognisance of this fact, different combinations of context features[31] have been used in training the classifiers, with the aim of establishing what sort of information is best suited to disambiguate the different types of ambiguities, as shown in table 3.19.

The study also seeks to investigate the effect of different context sizes on disambiguation accuracy for Swahili, and to find out if the standard two-word window applicable for other languages and especially English (Kaplan 1955), holds for Swahili. In this regard, different training data sets where the contextual information is obtained from a 2-, 10- and $>$10-word window[32] are prepared for each classifier.

The other research objective is to establish if performance would be improved significantly if a dependency-type grammar that grouped constituents into phrases e.g. Noun phrase were to be used, instead of the current constraint grammar parser that does not do

[31] The Noun prefix is excluded from the morphological features for WSD since it remains unchanged for all senses of an ambiguous word.

[32] The sentence-boundary restriction is applied for all cases, especially for the $>$10-window where it is more relevant.

Context Window size	Without NP Chunking	With NP Chunking
+/- 2	Data Set A	Data Set D
+/- 10	Data Set B	Data Set E
> 10	Data Set C	Data Set F

Table 3.20: Experiment data sets

so. A simple noun phrase chunker that groups noun modifiers together with the head noun into one phrase, was developed and applied to the SWATWOL output. The effect of this chunking is to increase the salient context for a noun. Different training data sets were then obtained for each of the classifiers using the newly tagged data. This research question arose from the observation that most nouns are succeeded by modifiers which provide only limited contextual information e.g. concordial prefixes, and for a small context window size e.g. 2, no selectional preference or co-occurrence information can be obtained. The example sentence 3.30 clearly illustrates this problem where the verb, which is the only source of predicate-argument information, is located some distance away from the head noun. The modifier for the head noun '*Kikosi*' comprises of seven words that precede the verb '*kilichosajiliwa*'. This is mainly due to the genitive construct that is employed extensively in Swahili expressing possessives, features (adjectives) etc.

Kikosi	*cha*	*sasa*	*cha*	*simba*	*chenye*	*chipukizi*	(3.30)
7/8-brigade	7/8-GenCon	now	7/8-GenCon	Simba	7/8-Poss-PRON	youngster	
wengi	*ki-li-cho-sajili-wa*	*na*	*Mwamwaja*	.....			
PL-Adj	7/8-Past-RelPRON-register-Passive	by	Mwamwaja				

"The current Simba brigade with many youngsters that was registered by Mwamwaja ..."

Therefore, for each of the 6 feature sets identified above, training data for each classifier was obtained from six different data sets as shown in table 3.20.

In addition to the above, a special data set was created that included only those feature variables that were determined to be important for semantic clustering using the SOM. This requirement affected only the Levin-based features where 131 features out of the total 183 features were selected. This was done with the aim of testing the effectiveness of using the SOM algorithm as a feature pre-selector for a supervised learning algorithm.

3.6.2 Training Bayesian Classifiers for Disambiguation

A BBN was constructed for each of the identified semantic classes. Each BBN, therefore, is a representation of what is typical of a particular class, and the resulting network structure provides an excellent opportunity to learn more about a particular semantic category, with respect to the contextual features used. It is worth noting however, that these networks are not unique for the given variables and their dependencies, and a network with a different topology could as well express more or less the same joint probability distribution. This is so since when the specified search time elapses, there may be hundreds of other equally probable networks given the data. This affects causal analysis where causal dependencies could provide an insight to the domain and the relationship between the domain variables. In this study, comparing network structures for different semantic classes could offer insight into the types of information that are useful in their discrimination. However, the BMT authors caution on the need for cautious interpretation of causal links. They attribute this to latent variables[33] which often induce sets of dependency statements, that cannot be described accurately by any Bayesian network, severely restricting the ability to automatically infer something about causalities, based only on statistical dependencies (Myllymäki et al. 2002). The BMT nonetheless provides tools to support naïve causal modelling assuming that there are no latent variables, and restricted latent variable causal modelling where latent variables are allowed, but with restricted dependency relationships. Causal analysis was out of the scope of this project, and was not done.

Examples of the network structures learned for the HUMAN classifier are shown in figures 3.11 - 3.14.

[33] A latent variable is one that for some reason has not been included in the data, and which has causal influence on the variables of the model

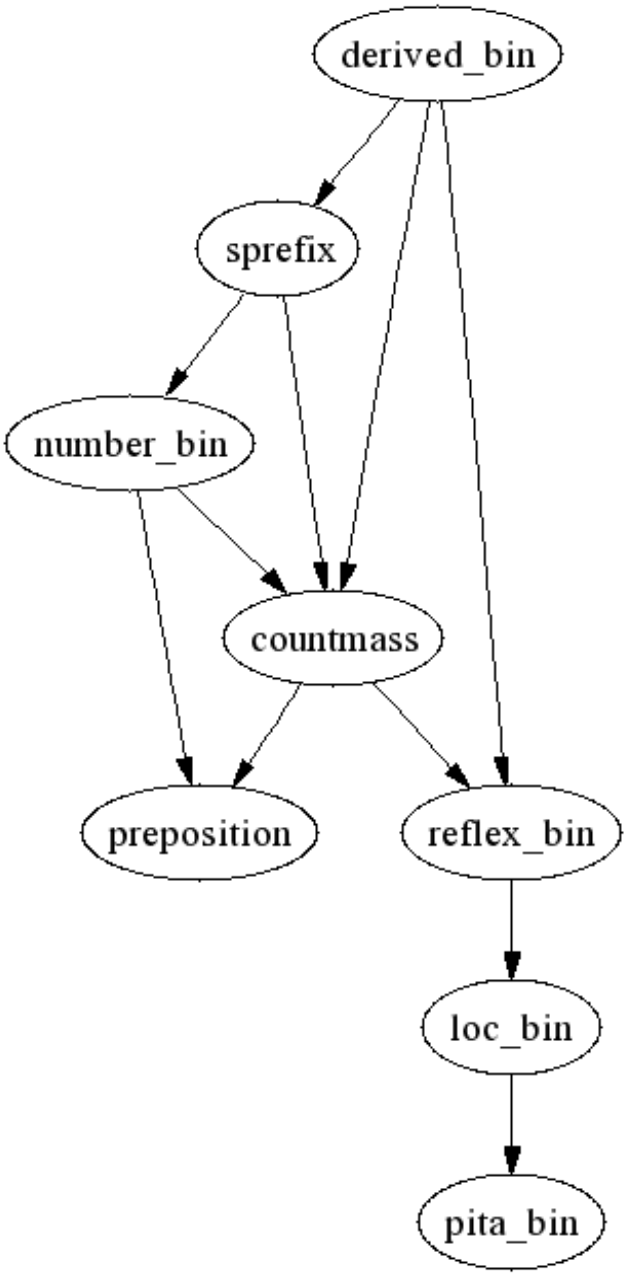

Figure 3.11: Human BBN (+/- 2; -NP Chunking)

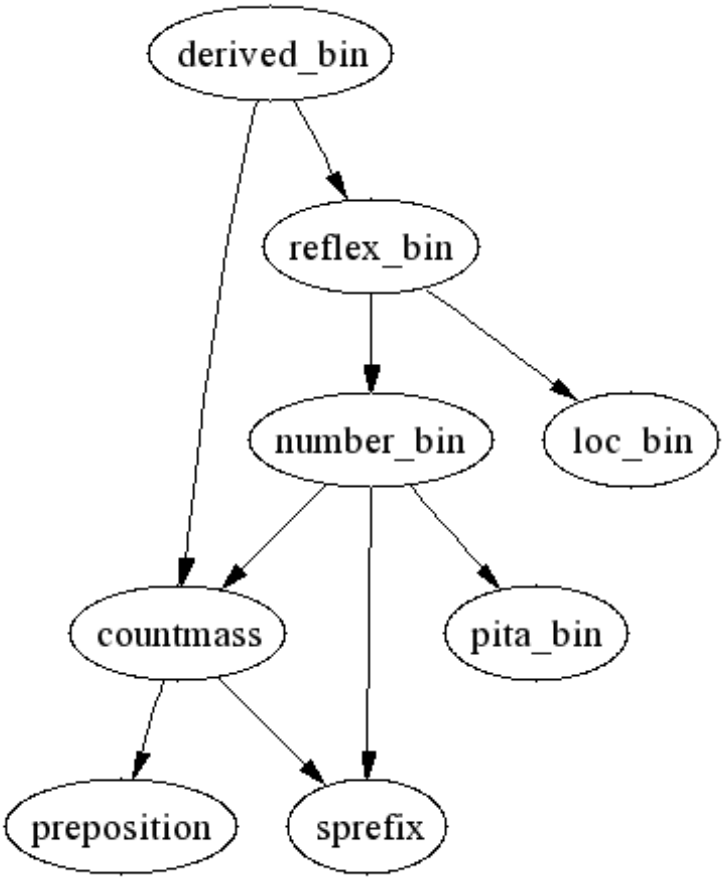

Figure 3.12: Human BBN (+/- 10; -NP Chunking)

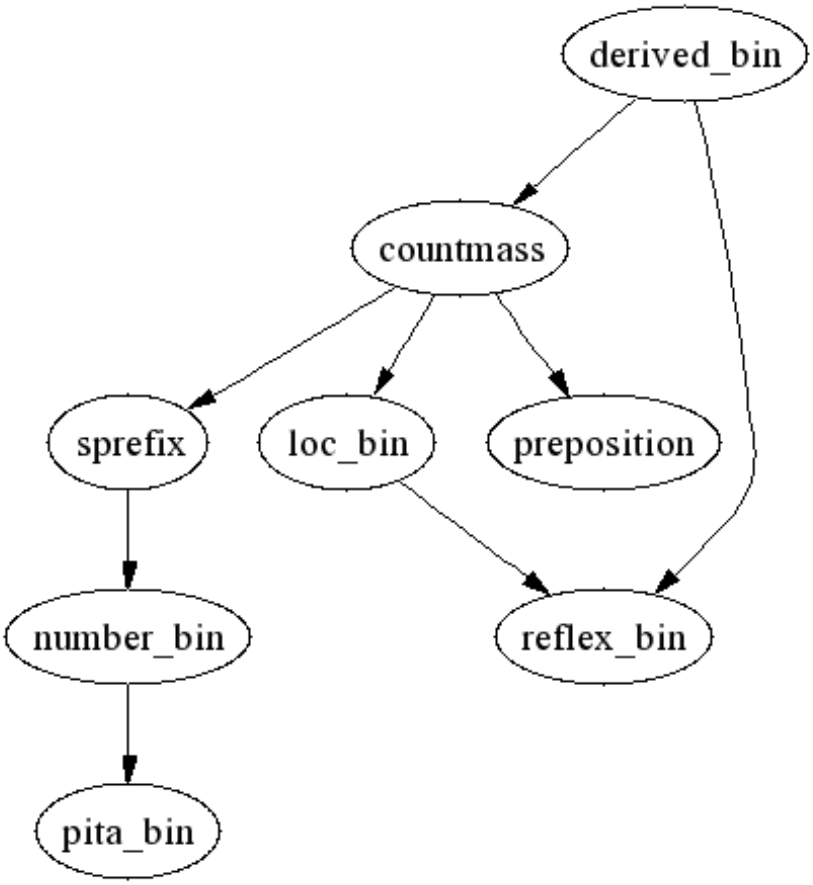

Figure 3.13: Human BBN (+/- >10; -NP Chunking)

The three sizes used for the context window have resulted in three different networks as shown in figures 3.11, 3.12 and 3.13, confirming that context window size is an important factor for disambiguation, as would be expected. The exact effect this has on disambiguation accuracy is discussed in the evaluation section in chapter 4.

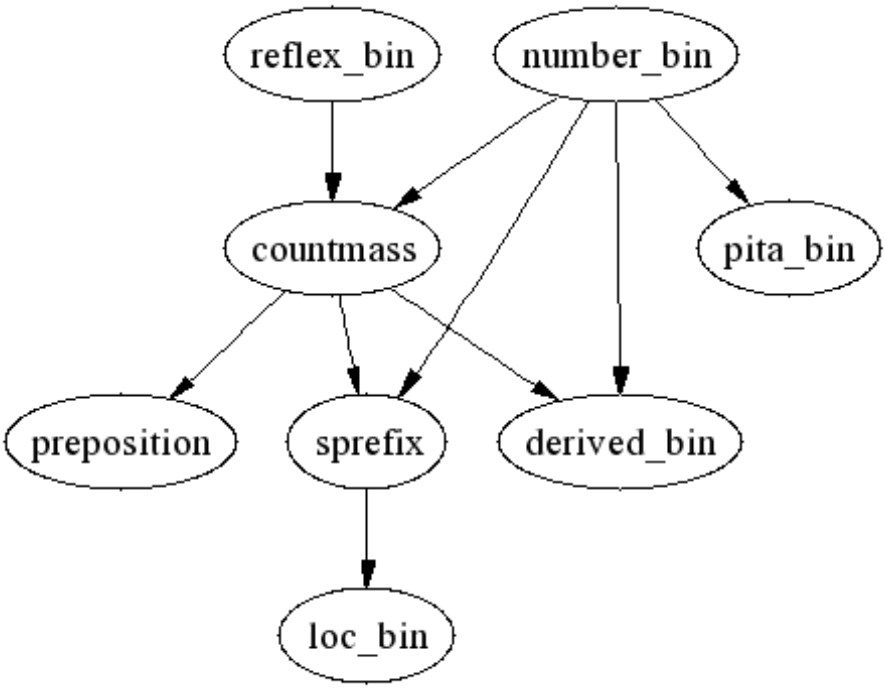

Figure 3.14: Human BBN (+/- 10; +NP Chunking)

Likewise, the difference in the structures shown in figures 3.12 and 3.14 is due to the different data sets used to obtain the training data. They have both been trained using feature set F, same context window size (10), the only difference being that for the latter, NP phrase chunking was performed. Results achieved when NP chunking has been done are compared to the default case (when no phrase chunking has been done) in chapter 4.

3.7 Summary

The methodology employed in developing the WSD solution has been presented in this chapter. As mentioned in chapter 2, a class-based approach has been adopted to address the data sparseness problem afflicting WSD research, and which is even more severe for less-studied languages that have limited linguistic resources. The semantic classes at the

core of the solution are determined empirically via unsupervised clustering using the SOM algorithm. The main motivation for this is to 'let the data speak', ensuring that only those classes, whose distinguishing semantic properties can be determined from Swahili textual data, are included in the solution. This is in contrast to approaches that rely on an external definition of classes say, a thesaurus or dictionary codes, which even though may result in a more refined set of classes capable of handling fine-grained senses, may suffer from lack of sufficient disambiguation information capable of supporting disambiguation of the ensuing fine meaning distinctions. A total of fifteen classes was identified for this study. With the classes in place, the next task is to analyze the lexical (translational) ambiguity inherent between Swahili and English, with respect to the derived classes. This step reveals the ambiguities that the WSD system should learn to disambiguate. The study focusses on two-way ambiguous nouns which comprise 72% of all ambiguous Swahili nouns, though the disambiguation methodology can handle n-way ambiguity, where n is the total number of semantic classes. The study exploits distributional clustering to automatically obtain labelled training data, from which 10% is reserved as test data. ML, and in particular BL, has been employed to learn probabilistic models that encode the linguistic nature of each of the classes, with respect to the contextual features chosen for this study. BBNs have been used to model the classes due to their ability to encode dependencies in the context features, a common characteristic of linguistic features. In addition, they cope well with missing and scarce data, a feature that complements the class-based approach to dealing with data sparseness. To address the central questions in WSD research regarding optimal context size and feature combination, as well as questions specific to Swahili NLP, the BBNs are trained on different data sets that test the performance of the WSD solution under varying conditions. The performance of the WSD system is presented in chapter 4.

Chapter 4
Evaluation

In this chapter, the evaluation of the Bayesian classifiers learned in chapter 3 is presented. The discussion proceeds with a summary of the resources required for testing and a definition of the evaluation metrics used in the study, as presented in sections 4.1 and 4.2 respectively. In section 4.3, the disambiguation results are presented. The results are reported for two sets of experiments:- set A, which is based on unambiguous nouns, serves as an approximation of the expected performance of the classifiers, while set B comprises a small set of hand-tagged ambiguous nouns that represent the main types of ambiguities identified in section 3.4.2. The latter is done with the aim of demonstrating the performance of the classifiers on an actual or real disambiguation task. A discussion on the obtained results is presented in section 4.4. The results are ordered to show the disambiguation performance for different classifiers in relation to the context window size, feature sets and noun phrase chunking, in tandem with the training parameters and conditions described in section 3.6.1. Section 4.4 discusses the achieved performance, paying particular attention to pertinent issues that arise in the ML paradigm adopted in this study, and their impact on the disambiguation performance. The testing configuration of the BBNs is altered to facilitate semantic tagging of unambiguous nouns, and the results obtained for this task are presented in the section 4.5.

4.1 Evaluation Resources

In the ideal setup, formal evaluation of a WSD system would require a sizeable hand-annotated test corpus containing several ambiguous words that would provide a gold standard for evaluation. In addition, performance figures for other systems on the same task and evaluated against the same gold standard would be required in order to benchmark the performance of the developed system. However, in reality, this ideal is rarely met and less so for the specific task of this study - Swahili WSD.

Rather than embark on the costly exercise of obtaining an extensive gold standard that covers all the different ambiguity types identified in section 3.4.2, the developed system is evaluated on the related task of semantic class categorization for unambiguous nouns. In addition to alleviating the need for a sizeable, elaborate gold standard, this approach allows extensive testing of the system on different ambiguity types using many different nouns and in varying test conditions, than would be possible if relying only on a small hand-tagged test corpus. Nonetheless, a small gold standard for a handful of ambiguous nouns is created for the purpose of validating the performance figures obtained using unambiguous nouns, as well as bringing to the fore important issues pertaining to the training of probabilistic classifiers for WSD. The criteria for choosing these words, as well as the relevant statistics on the resulting test corpus, are presented in section 4.3.2

In the absence of comparative performance figures for other systems on Swahili noun WSD, a glass box evaluation approach is adopted where different system aspects and components and their significance on the obtained performance are discussed. Where relevant, these are contrasted to those of comparable systems based either on similarity of task, statistical NLP approach or disambiguation information acquisition and resource requirements.

4.2 Evaluation Metrics

Common metrics in WSD evaluation have been used to quantify the performance of the developed system. These are:

$$\begin{aligned}
\text{Precision (P)} &= \frac{TP}{TP+FP} \\
\text{Recall (R)} &= \frac{TP}{TP+FN} \\
\text{F}_1\text{ Measure} &= \frac{2 \times P \times R}{P+R} \\
\text{Accuracy (Acc)} &= \frac{TP+TN}{P_t+N_t}
\end{aligned}$$

where TP, TN, FP and FN refer to true positives, true negatives, false positives and false negatives respectively (as classified by the system) and P_t and N_t refer to the total

number of positive and negative examples in the test set respectively. In a binary class-based classification context, the terms positive and negative as used in these definitions are associated with membership to one of the two semantic classes involved in the classification (senses). For example, where disambiguation involves the classes HUMAN and ANIMAL, P_t and N_t refer to the total number of test occurrences belonging to class HUMAN and ANIMAL respectively, while TP (TN) refers to the HUMAN (ANIMAL) test occurrences correctly classified as such by the system. Likewise, FP (FN) refers to those ANIMAL (HUMAN) test occurrences that have been misclassified by the system as belonging to class HUMAN (ANIMAL).

Due to the performance trade-off between precision and recall, the F_1 measure, computed as a harmonic mean between these two values, yields a single number by which performance can be measured. This provides a convenient way to compare the performance of two or more classifiers on the same problem, ranking them in order of quality of prediction. In this study, the F_1 measure is used, with equal weight assigned to both precision and recall.

Accuracy is a commonly used and straightforward metric which simply reports the percentage of correct classifications. The accuracy value enables comparison of a classifier's performance against a given base line such as the majority classifier which acts as the lower bound for the performance of probabilistic classifiers. The majority classifier simply selects the MFS as the correct sense for an ambiguous word. In this study, the BBN classifiers have also been rated against the simpler NB classifiers.

Manning & Schütze (1999) note that the F_1 measure and accuracy are different objective functions with accuracy being sensitive only to classification errors, while the F_1 measure, by definition, is more sensitive to type I and II errors (FN and FP). Consequently, the F_1 measure prefers results with more true positives.

4.3 Results

Performance of the WSD classifiers is presented in this section where the evaluation is based on the metrics presented in the previous section. As explained in chapter 3, Bayesian classifiers were induced from training data for each of the fifteen semantic classes listed in table 3.7. Each BBN is thus a representation of the typical linguistic form of a given semantic category, as defined by the context features used in its formulation. To disambiguate any of the ambiguity types listed in section 3.4.2, the pair of classifiers for each of the involved classes is used to predict the probability that the given test vector belongs to one of the two classes. The winning classifier, and hence the most probable sense given the current context, is chosen as that which awards a higher probability i.e. if $\mathbf{H}$ and $\mathbf{A}$ represent the BBN classifiers for class HUMAN and ANIMAL respectively, then the test data vector, d_i, is classified as belonging to class (sense) HUMAN if $P(d_i|\mathbf{H}) > P(d_i|\mathbf{A})$.

The performance of the 15 learned BBNs in disambiguating the major ambiguity types important for Swahili-English MT was estimated by testing the classifiers on the similar task of semantic category classification for test vectors obtained from unambiguous nouns and whose true classes are therefore known. The overall results for test set A, obtained by averaging the performance achieved over all the different ambiguity types, are presented in section 4.3.1. In this section, evaluation conditions relating to optimal context window size, overall best feature set and the effects of surface chunking of noun phrases are addressed. The overall results are an averaged account of the general performance of the WSD solution. Test set B results obtained by testing the system on a small set of hand-coded ambiguous words are presented in section 4.3.2.

4.3.1 Set A: Unambiguous Nouns - Overall Performance

Using the F_1 measure, the performance results displayed in figures 4.15 and 4.16 show that regardless of the feature set or context window size, the BBN classifiers, both BBN-ac and BBN-sc[34], consistently outperform the NB classifiers, with an average improvement

[34] For BBN-ac (all components) all the 207 feature components were included in the BBN topology, while

of 7.1 percentage points and 8.3 percentage points respectively. It is worth noting that the WordNet-based features (**W**) yielded the smallest improvement (2 percentage points), while the largest gain (12 percentage points) was obtained using Levin-based features (**L**). This confirms what is intuitively expected since the fine granularity of Levin's verb classes (183 classes) results in highly correlated (dependent) features. In contrast, WordNet has only 15 verb classes, and since these are very coarse grained, there is not as much correlation between the different classes compared to Levin's classes. For example, while WordNet has one class for **consumption** verbs, Levin has 7 classes for the same (**eat**, **chew**, **gobble**, **devour**, **dine**, **gorge** and **feed**). For those cases where Swahili does not match such fine granularity, the same Swahili verb occurs in several classes. For example, while **chew**, **gobble**, **devour** and **gorge** have different Swahili translations, **eat**, **dine** and **feed** are all translated as '*la*', making these classes highly correlated. The NB's independence assumption is thus severely violated for this feature set compared to the WordNet-based one. In contrast, the BBN takes into account these dependency relations, making it a better model for the feature set and thus yielding a much better improvement over the NB results for feature set **L** compared to **W**.

As presented in section 3.6.2, the number of Levin-based features was reduced from 183 to 131 by using the SOM as a feature selector. The results achieved by the BBNs trained using the SOM-selected components for this feature set are an improvement over those where all the 183 features were included, registering an increase of 7 points for the F_1 measure.

for BBN-sc (SOM components), only 155 components that were important for semantic clustering using the SOM algorithm were included.

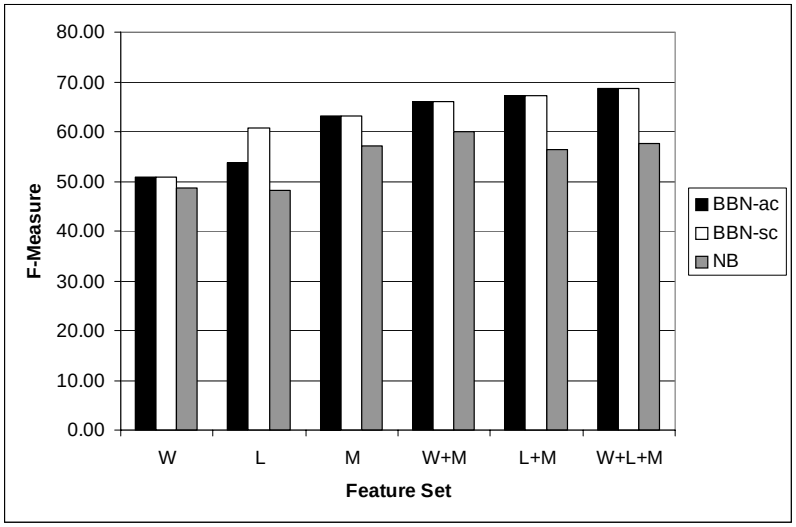

Figure 4.15: Performance based on different feature sets (**W**ordNet, **L**evin, **M**orph. + POS)

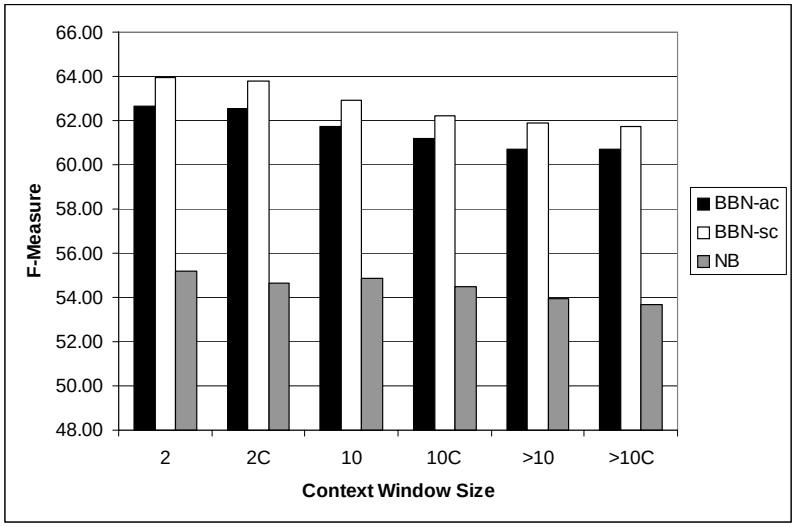

Figure 4.16: Effect of NP-chunking (C) and varying the context window size on performance

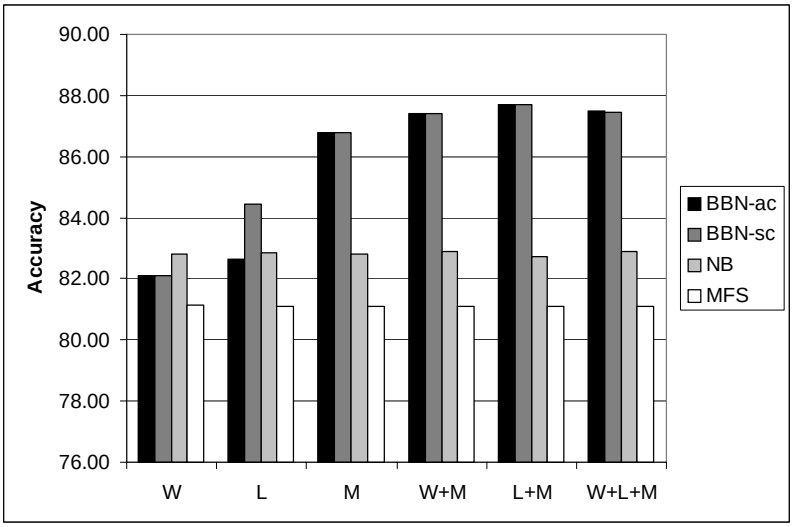

Figure 4.17: Accuracy of BBN classifiers compared to baseline classifiers: effect of different feature sets

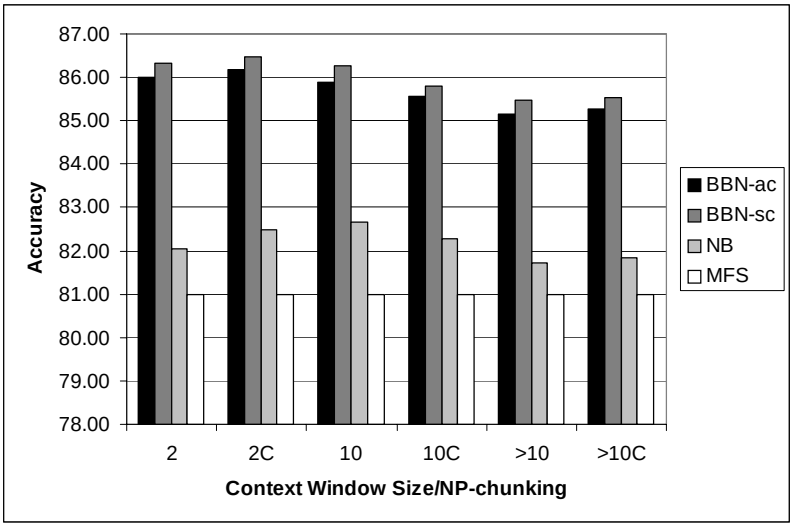

Figure 4.18: Accuracy of BBN classifiers compared to baseline classifiers: effect of varying Context window size/NP-chunking

Despite the high average for the MFS (81%), both the BBNs and NB classifiers manage to improve on the accuracy of the majority classifier (MFS) for all feature sets, with an

average of 5 points and 1.2 points respectively (fig. 4.17). Also, as with the F_1 measure, the BBNs achieve a higher accuracy of 4 points over the NB classifiers, as does BBN-sc over BBN-ac (2 pts) for the Levin-based feature set.

Figures 4.15 - 4.18 summarize the performance of the developed WSD solution under various test conditions relating to feature set, context window size and NP-chunking, based on BBN-sc. Considering individual feature sets, morphology + PoS features (**M**) yield the best results as shown in figures 4.15 and 4.17, compared to verbal-based feature sets (**W** and **L**). Feature set **M** registers the highest F_1 measure of 63.1 and 86.8% accuracy, compared to 60.6 and 50.9 (F_1 measure) and accuracies of 84.4% and 82.1%, for **L** and **W** respectively. The highest overall F_1 measure of 69 is obtained by a combination of all three feature sets (**W+L+M**), while the highest overall accuracy score (87.7%) is achieved by feature set **L+M**, that combines Levin-based features with the Morphology + PoS features. This difference in the best performing feature set can be attributed to the objective differences between accuracy and F_1 measure as explained in section 4.2. Of the three, feature set **W** achieves the lowest performance for both measures, an indication that it is the weakest set in terms of discriminatory power (affecting classification/discrimination between classes) and in representing the typical semantic and linguistic element of a given class (affecting positive identification of a class). The F_1 measure, which seeks to maximize positive identification (TP and TN), favours feature set **W+L+M** which exploits the complementary and redundant information contained in the three feature sets. Consequently, the best score is achieved using this set. In contrast, accuracy, which seeks to minimize classification errors, would necessarily benefit from a feature set with more discriminatory power, and thus feature combination **L+M** which excludes **W** obtains the best accuracy result.

With regard to the optimal context window size for WSD based on local context features, figures 4.16 and 4.18 show that a small window of two words on either side of the ambiguous word is sufficient for extracting useful disambiguation information, with higher (though only slightly) overall F_1 measures and accuracy figures obtained for window size 2 compared to size 10 or greater. This is an important empirical validation of Kaplan's (1955)

Context Window Size	Without NP-chunking	With NP-chunking
2	57%	71%
10	94%	99%
>10	99.01%	99.98%

Table 4.21: Percentage of training contexts containing a verb within the specified window

observation that two is an optimal context window size for sense resolution, as explained in section 2.2.1, for the case of Swahili data.

NP chunking was done with the aim of yielding a compact context that contains more sources of potentially useful disambiguation information than would have been available otherwise. This is especially important when considering feature sets that exploit grammatical relations such as the selectional preferences based on WordNet and Levin verb classes that have been used in this study, with the aim of linking the target noun to the head verb in the sentence, and deriving semantic information from the ensuing grammatical relationship.

As shown in table 4.21, for all context window sizes, NP-chunking increases the probability of including a verb within the target noun's context. This results in an increase in the context's saliency with respect to selectional preference information. NP-chunking is especially relevant for the smallest window size (2). As would be expected, for both sets of experiments based on verbal features (**L** and **W**), the results obtained using the chunked contexts are slightly better than those where NP-chunking was not done. These results are shown in figures 4.19 and 4.20 respectively. However, as figure 4.21 illustrates, chunking does not improve performance for morphology + PoS features, as the majority of these are already available from context 0 (the target noun itself) and 1 (target noun's immediate modifiers), in the case of Swahili[35], and as such no major benefit is gained from chunking. An interesting observation is that NP-chunking actually depreciates the performance for this feature set. This phenomenon is explained in detail and illustrated with an example in section 4.4.

[35] Out of the 8 feature types in set **M**, 3 are obtained from the target noun itself (locational suffix, plural/singular prefix, derivational suffixes), while 4 can be obtained from position +/- 1 (subject prefix, number, preposition), while only 1 is collected from the head verb (reflexive marker).

From the results displayed in figures 4.15-4.21, the optimal experimental conditions with respect to feature set, context window size and NP-chunking that yield the best overall WSD performance are shown to be morphological + PoS information extracted from a small context window of +/- 2, and without the need for NP-chunking. Validation of these hypotheses based on disambiguation of a small set of ambiguous nouns is presented in section 4.3.2.

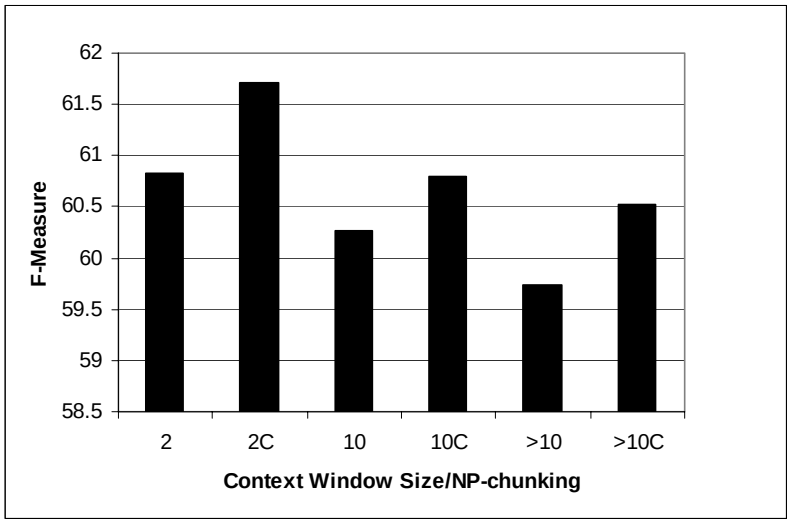

Figure 4.19: Effect of NP-chunking on performance: Levin-based features

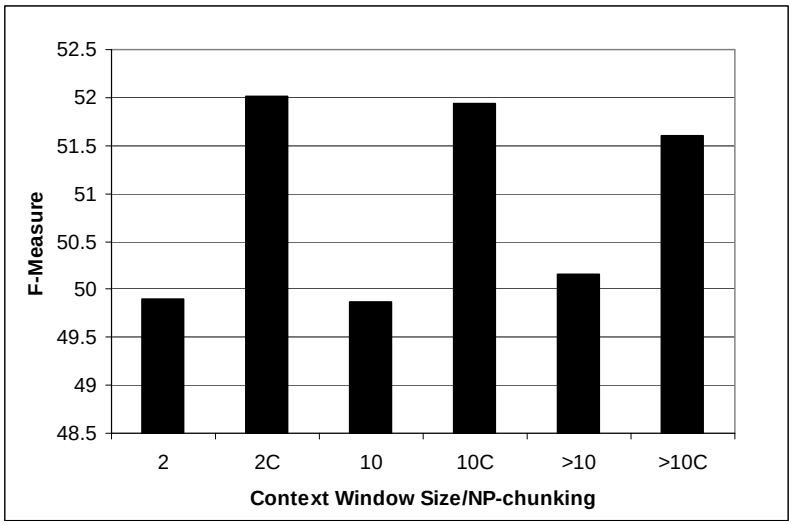

Figure 4.20: Effect of NP-chunking on performance: WordNet-based features

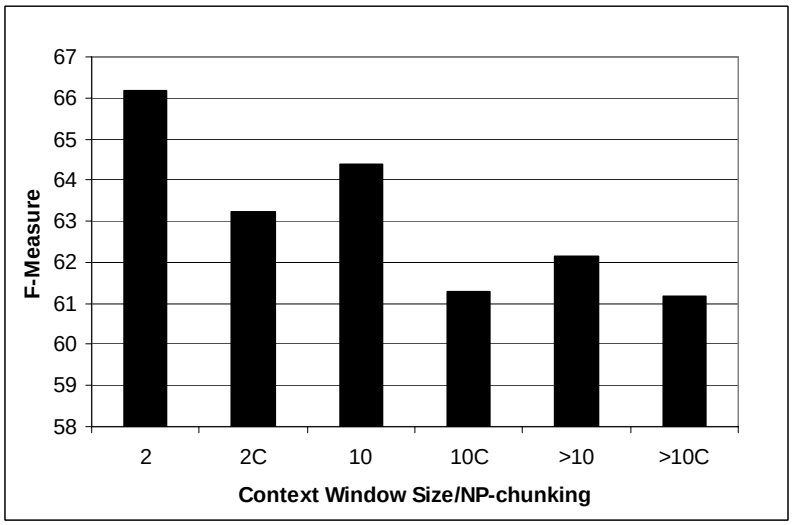

Figure 4.21: Effect of NP-chunking on performance: Morphological + POS features

4.3.2 Set B: Ambiguous Nouns

In this section, the performance of the BBN classifiers is tested on ambiguous nouns. The test set was carefully chosen ensuring that: a) some of the most common ambiguities relevant for Swahili-English MT identified in section 3.4.2 are represented; b) the senses of the selected nouns cover the main types of semantic ambiguity, i.e. homographs, metonyms and metaphors and c) examples for both senses of each word can be obtained from the Swahili corpus. A total of 16 ambiguous nouns, involving 11 semantic classes, were selected following these criteria, and are shown in table 4.22.

Testing the WSD system on actual ambiguous nouns not only demonstrates the performance of the learned classifiers on a real disambiguation task, but more importantly, highlights important issues that should be considered when porting WSD systems. Of critical importance is the role of bias, i.e. the distribution of the number of examples per sense in the training and test data, on the performance of probabilistic classifiers. Agirre & Martinez (2000) have reported that results degrade significantly when the training and testing samples have different distributions for the senses. For test set B, the bias factor is important since the sense distributions in the training data are linked to their corresponding semantic class sizes, which have been estimated from the number of occurrences in the Swahili corpus, of member (unambiguous) nouns. Clearly, this is a very rough estimate and would, for the most part, not be consistent with the actual distribution of senses of individual ambiguous nouns. The differences in the sense distributions between the automatically-acquired training corpus and the actual distribution as determined from the hand-tagged test corpus for each of the 16 nouns are shown in table 4.22. As shown, a few of the words have a comparable distribution, e.g. '*mkunga*' and '*tembe*', while the rest differ significantly, with words such as '*nyanya*' and '*sindano*' having completely opposing training and test sense distributions.

In light of these differences in sense distribution, different bias settings[36] were used during disambiguation of the test set, with a view to determine how disambiguation accuracy is affected under each of these settings. The settings are: a) automatic sense distrib-

[36] The bias settings are applied to the training data.

		Auto	Bias	Test	Bias
Noun	**Classes (senses)**	**C1**	**C2**	**C1**	**C2**
juma	HUMAN-TIME	59.6	40.4	80.3	19.7
mkunga	HUMAN-ANIMAL	95.6	4.4	97.1	2.9
kiongozi	HUMAN-ARTIFACT	79.7	20.3	97.8	2.2
kirukanjia	HUMAN-ANIMAL	95.6	4.4	33.3	66.7
nyanya	HUMAN-FOOD	90.1	9.9	6.3	93.7
korongo	ANIMAL-LOCATION	7.0	93.0	68.4	31.6
ndege	ANIMAL-VEHICLE	35.1	64.9	72.6	27.4
buibui	ANIMAL-DRESS	52.0	48.0	43.3	56.7
tembe	LOCATION-FOOD	84.7	15.3	83.3	16.7
ua	LOCATION-PLANT	98.9	1.1	63.8	36.2
mwezi	LOCATION-TIME	47.3	52.7	19.2	70.8
pango	LOCATION-MONEY	77.9	22.1	70.3	29.7
sindano	ABSTRACT-ARTIFACT	87.3	12.7	28.2	71.8
saa	ARTIFACT-TIME	27.3	72.7	5.6	94.4
bakora	ABSTRACT-ARTIFACT	87.3	12.7	40.5	59.5
usukani	ABSTRACT-TIME	87.3	12.7	47.1	52.9

Table 4.22: Test Nouns: sense distribution in automatically-acquired training corpus vs. hand-tagged test set

ution, determined as the class size of the corresponding semantic classes for each word's senses; b) no bias, where an equal amount of examples was used for each semantic class and c) test set bias, which was determined from the small hand-tagged test corpus, and which represents the true[37] sense bias for each word. For setting a) no retraining of BBNs was done and the same BBNs used for set A nouns were used to disambiguate the test set. For bias setting b) and c), new classifiers were trained with data that reflects the required bias settings, and then used for test set disambiguation. The results are shown in table 4.23[38].

By looking at the results based on the automatically-acquired sense distribution (columns 3 and 4), the important role of bias is clearly evident, with results higher than the MFS[39]

[37] In this case, the true bias is taken as that determined from the hand-tagged examples for each word, retrieved from the Swahili corpus.

[38] The accuracy results shown in columns 4, 6 and 8 represent the best possible result for each word, regardless of the feature set, while the feature set column (col. 9), gives the feature set that consistently yields the best result for a given word, under varying bias settings. The accuracy columns therefore indicate the best performance for a word, while the feature set column shows the best average performance for a word, indicated by the corresponding feature set.

[39] Values are shown in bold face where the accuracy obtained is higher or equal to the MFS baseline.

		Auto		None		Test		
Noun	**Type**	**MFS**	**ACC.**	**MFS**	**ACC.**	**MFS**	**ACC.**	**Feature Set**
juma	Homo	59.6	**84.0**	50	**71.9**	80.3	**83.8**	W+L+M
mkunga	Homo	95.6	**97.1**	50	**94.1**	97.1	**97.1**	M
kiongozi	Meta	79.7	**97.8**	50	**97.8**	97.8	**97.8**	M,L+M
kirukanjia	Meto	95.6	33.3	50	**66.7**	66.7	**66.7**	W,L,M,L+M
nyanya	Homo	90.1	12.5	50	**93.8**	93.7	**93.8**	M
korongo	Homo	93.0	57.9	50	**84.2**	68.4	**84.2**	M
ndege	Meta	64.9	**81.2**	50	**80.1**	72.6	**82.3**	M
buibui	Homo	52.0	**66.7**	50	**70.0**	56.7	**66.7**	L+M
tembe	Homo	84.7	**100**	50	**100**	83.3	**83.3**	L+M
ua	Homo	98.9	63.8	50	**63.8**	63.8	**74.5**	L+M
mwezi	Meta	52.7	**85.9**	50	**88.5**	70.8	**82.1**	L
pango	Homo	77.9	**78.2**	50	**75.2**	70.3	**79.2**	L+M
sindano	Meta	87.3	28.2	50	**71.8**	71.8	**76.9**	L
saa	Meto	72.7	**91.6**	50	**86.9**	94.4	**94.4**	L
bakora	Meto	87.3	40.5	50	**54.1**	59.5	**59.5**	W
usukani	Meta	87.3	70.6	50	**64.7**	52.9	**64.7**	M
AVERAGE		**80.0**	**68.1**	**50.0**	**79.0**	**75.0**	**80.4**	

Table 4.23: Disambiguation accuracy obtained using varying sense biases

baseline being obtained for all nouns where the sense distribution is consistent in both the training and test data. In contrast, for those nouns with an opposing bias, performance better than MFS is only achieved for two of them, '*ndege*' and '*buibui*'. When no bias is used in training, the MFS baseline is exceeded for all words. The same is true when the test set bias is used in training. However, as the average accuracy for all the words shows, using no bias achieves performance that is comparable to that achieved using the test set bias.

From the results, the semantic ambiguity type exhibited by a noun's senses does not seem to be an important factor in the disambiguation accuracy, with good performance achieved for homographs, metonyms and metaphors. Due to the class-based approach adopted in the study, disambiguation performance is determined more by the specific semantic classes that represent the noun's senses, and the feature set used, with different feature sets being better discriminators between different pairs of classes, irrespective of the semantic ambiguity type. For example, feature set **M** seems to be more applicable when one of the involved classes is animate e.g. '*mkunga*', '*nyanya*', '*korongo*' and '*ndege*'. This is due to the semantic importance of the animate prefix which is uniquely associated with humans and animals. Selectional preference information is vital for the other cases where

such a dominant distinguishing feature is absent (‘*saa*’, ‘*sindano*’ and ‘*mwezi*’). Just as was the case for test set A, for all words, feature set **M** achieves the best average disambiguation accuracy (72.8%), followed by **L** (69.5%) and lastly by **W** (68%).

4.4 Analysis

The disambiguation results achieved using the BBN classifiers have been presented in the preceding sections. In this section, a discussion of these results with respect to the training conditions presented in section 3.6.1 is presented. A careful analysis of the achieved results is given, with examples drawn from the disambiguation of ambiguous nouns to highlight the main causes of erroneous classification.

Feature Occurrence Two types of contextual information were used in the experiments - morphology + PoS information and grammatical relations (selectional preferences). For the latter, two sources of verbal semantic information were used - WordNet and Levin. The results presented in section 4.3.1 show that Morphology + PoS features achieved the best overall performance compared to verbal-based features. In addition to the important semantic information relayed by Swahili’s morphology (see section 3.3.1), the fact that most of this information is contained within the target noun and its immediate modifiers (context position +/-1) makes this feature set very dominant in the training data, as this information is always available, unlike the verbal-based features where the verb may not always be found within the context window. Given the probabilistic nature of the ML paradigm in use, the frequency of occurrence for any given feature in the training set has important consequences for all the subsequent probabilities that will be awarded to it. Hence, morphological features on their own achieve almost the same performance as that achieved when combined with either or both of the other feature sets (**W** and **L**). In addition, feature set **M** comprises mostly of closed class items which have fixed meanings and not being too numerous, generally makes them good, reliable features. The coverage of the edited[40] bilingual dictionary was not complete and as a

[40] Swahili verbs in the bi-lingual dictionary were manually edited to enable a higher number of hits when

result, the number of training contexts for which WordNet and Levin tags could be obtained for occurring verbs, was 88% and 74% respectively, contributing to the poorer performance achieved using feature sets **W** and **L**. In addition, the performance of these feature sets is adversely affected by the assumption that the first sense listed for polysemous verbs is always the correct translation of the Swahili verb in question. Though no thorough statistics have been made regarding this assumption, there are occasions when it does not hold. A consequence of this violation is that a Swahili noun is associated with wrong verbal semantics and this contributes to its misclassification. For example, the verb '*ajiri*' is unambiguous meaning *employ (hire,engage,retain)*. However, this verb is ambiguous in English with the first listed sense in WordNet being *use or* utilize (**consumption**), and the second *hire* or *engage* (**social**). In this case, rather than associate the subject or object of the verb '*ajiri*' with the preference semantics of a **social** verb, these are associated erroneously with a **consumption** verb, where for instance a HUMAN noun is taken to be a legal object of a **consumption** verb rather than a **social** verb.

Feature Set The results also show that the Levin-based feature set achieves a better overall performance than that based on WordNet. The finer granularity of Levin's classes allows for a much finer distinction between semantically-close classes than is achievable using WordNet's general classes. Consequently, some of Levin's verb classes can only be associated with a particular semantic class and therefore serve as unique identifiers for that class. For example, WordNet's **communication** verb class is realised using several 'specialised' Levin classes (**message_transfer**, **speakmanner**, **talk**, **chitchat**, **say**, **communication_instrument**, **complain**, **advise and animal_sounds**). Clearly, Levin's **animal_sounds** class is sufficient to distinguish between classes HUMAN and ANIMAL, something that WordNet's **communication** cannot achieve. Other examples of specialised Levin classes include **vehicle** and **drive** verbs which are important discrimi-

querying WordNet and Levin's classes. For example, the dictionary entry for the Swahili verb *egama* is 'be in a leaning, resting or reclining position', and had to be edited to 'lean, rest or recline' in order to obtain the corresponding WordNet tag successfully.

nators for class VEHICLE, unlike WordNet's more general **motion** class which does not discriminate animate versus inanimate types of motion. For this reason, Levin-based selectional preferences are more effective at disambiguation compared to WordNet-based ones, despite the fewer number of training contexts where the corresponding Levin tag was successfully obtained.

NP-chunking As shown in section 4.3.1, NP-chunking was beneficial in obtaining selectional preference information from contextual verbs, but resulted in deteriorated performance for the morphological + PoS feature set. The problem stems from the unordered context where, due to a lack of syntax parsing or phrase chunking for all sentence constituents, the verb within an ambiguous noun's context does not always have a grammatical relationship with it i.e. the noun may not always be the subject or object of the verb. As a result, morphological information contained in the verb, specifically the subject prefix, may not be consistent with the semantics of the noun sense in question, and this results in erroneous classification. For example, sentence 4.31 contains the target word '*juma*' which is ambiguous between a *proper noun* (HUMAN) and *week* (TIME).

naye alipofufuka	alfajiri siku ya kwanza ya	juma	,	alimtokea kwanza	(4.31)
and when he resurrected	in the morning of the first day of	week/name	,	he appeared first to	
mariamu magdalena	ambaye kwamba alimtoa pepo saba				
mary magdalene	from whom he cast out seven demons				

The analysis obtained from SWATWOL (considering a +/-2 word context) is shown in figure 4.22[41]. From this context, crucial morphological information contained after the noun is obtained from the subject prefix of the verb '*alimtokea*'. However, from the original sentence the subject of this verb is the sentence-initial pronoun '*naye*'. Since 1/2-SG3-SP is an animate (human) prefix, this occurrence of '*juma*', which in this sentence refers to the TIME sense (*week*), is erroneously classified as HUMAN. Given the disambiguation improvement NP-chunking makes for verbal-based features and the deterioration suffered using morphology + PoS features due to lack of comprehensive

[41] Currently, SWATWOL tags all occurrences of the word '*juma*' with the TIME sense (week), and thus disambiguation for this word is necessary to determine those instances where it is used in the HUMAN sense.

chunking, it can be concluded that complete phrase-chunking or dependency syntax parsing for Swahili is a worthwhile endeavour needed to improve various NLP applications, such as the WSD task undertaken in this study.

```
"<kwanza>" "kwanza" NUM NUM-INFL ORD { first }

"<ya>" "ya" GEN-CON 9/10-SG

"<juma>" "juma" N 5a/6-SG { week } AR

"<,>" "," COMMA

"<alimtokea>" "tokea" V 1/2-SG3-SP VFIN PAST 1/2-SG3-OBJ OBJ { put out ,
remove , publish , produce , generate , offer to someone , subtract , reduce } SVO
EXT: STAT APPL :EXT
```

Figure 4.22: SWATWOL analysis for disambiguation context - ‘*juma*’

SWATWOL analyses The disambiguation contexts are analysed using SWATWOL and consequently the quality of the analyses impacts on the classifiers’ performance. In general, SWATWOL achieves very high accuracy in morphological and part of speech tagging. There were a few cases, however, where morphological disambiguation fails yielding a wrong analysis, which in turn results in sense misclassification. For example, sentence 4.32 contains the target word ‘*jini*’ with translation equivalents *genie* (HUMAN) or *gin* (FOOD). Figure 4.23 shows the corresponding SWATWOL analysis when considering a +/- 2 context window. The word ‘*wala*’ is ambiguous with respect to part of speech. While in this sentence it refers to the conjunction *nor*, SWATWOL analysis gives a verbal interpretation yielding the verb ‘*la*’(*eat*). Since ‘*jini*’ is ambiguous between classes HUMAN and FOOD, it is misclassified as *gin* in this instance as it occurs

as the direct object of an **eat** or **consumption** verb.

ulikuwa	utoto	na	utoto	hauna	hatari	ya	uke	wala	uume	(4.32)
it-was	childishness	and	childishness	has-not	danger	of	feminity	nor	masculinity	
,	shetani	wala	jini	,	machaka	wala	misitu			
,	devil	nor	genie	,	thickets	nor	forests			

```
"<shetani>" "shetani" N 9/6-0-SG { satan , devil , demon , wicked person } AR HUM
"<wala>" "la" V 1/2-SG2-SP VFIN PR:a { eat } SV SVO MONOSLB
"<jini>" "jini" N 9/6-0-SG { genie , sprit , demon , Belial , wicked person } AR HUM
"<,>" "," COMMA
"<machaka>" "chaka" N 5a/6-PL { clump of trees , thicket }
```

Figure 4.23: SWATWOL analysis of disambiguation context - '*jini*'

Contextual Information According to Weiss & Kulikowski (1991), classification performance is more dependent on the training data and feature set than on the individual ML algorithm. The choice of the feature set is of critical importance to the predictive ability of the learned classifier. In this study, dependency (relational) features comprising of overtly-marked morphological and part of speech features were used in conjunction with selectional preferences derived from simple grammatical relations between the target noun and contextual verbs. These features represent only local context, while global or domain context has been left untapped. One reason for this is that the WSD method developed is targeted at general WSD where a wider coverage of words using the same basic classifiers is achieved, in contrast to word-specific WSD where classifiers are constructed for each individual target word. For the latter, identification and inclusion of global context in the form of collocations and word co-occurrences is straightforward. This type of disambiguation information has been shown to be very useful for WSD, since words tend to have only one sense for a given discourse or collocation (Gale et al. 1992*b*), (Yarowsky 1993). In contrast, due to the need to use only that disambiguation information which is applicable to a range of different words, the class-

based approach adopted in this study suffers from information loss especially of topical information and collocations, which are specific to individual words. For example, for the ambiguous word '*nyanya*', co-occurrence information is usually sufficient for disambiguation, with the HUMAN reading applicable in example 4.33a, and the FOOD reading in example 4.33b, regardless of the other information carried in the relational features.

$$\underset{\text{grandfather}}{\text{Babu}} \quad \underset{\text{and}}{\text{na}} \quad \underset{\text{grandmother}}{\text{nyanya}} \tag{4.33a}$$

$$\underset{\text{onions}}{\text{Vitunguu}} \quad \underset{\text{and}}{\text{na}} \quad \underset{\text{tomatoes}}{\text{nyanya}} \tag{4.33b}$$

The case for using collocations and other forms of global context to supplement local context is evident in cases where the local features are not observed in a given context. In such cases, a system that considers a wider (topical) context will have some other information that could be useful to guide disambiguation, rather than just defaulting to the MFS, as is the case in this study. Another alternative solution in such cases would be to include a rule-based disambiguation system to provide complementary disambiguation information (Hurskainen 2004*b*).

Real World Knowledge However, even if all the information useful for disambiguation that exists in textual data was extracted and incorporated into the training of an automatic classifier, there are numerous cases where disambiguation fails, due to the absence of critical extra-linguistic or real world knowledge. Numerous psycholinguistic studies have shown that human beings rely on world knowledge and inference, in addition to local context, domain knowledge and frequency data (Liddy 1998), when disambiguating word senses. Incorporating world knowledge into disambiguation systems has proved challenging, and even though some effort towards using ontologies and semantic webs to supply this information has been undertaken (Ciaramita et al. 2003), providing machines with this knowledge and equipping them with mechanisms that allow them to reason and infer meaning from it, has proved to be a difficult task. Therefore, examples such as those shown in 4.34 and 4.35 are still beyond the disambiguation scope of automatic classifiers that rely only on textual knowledge for disambiguation. In example

4.34, choosing the right translation of '*pau*' (*rafters)*, requires a deep understanding of the meaning of the different words as well as their compositional meaning. If a person is not familiar with this idiomatic expression, the type of world knowledge required to correctly choose the right reading includes: knowing that rafters are elements of a house's roof; roofs are located at the top of a house structure; humans sleep in houses (usually) and that when someone sleeps on their back, they are facing upwards to the roof. Likewise, for example 4.35, the senses of '*sindano*' have a metonymic relationship where the ambiguity is between the instrument (*needle*) and the act of using the instrument (*injection*). In this case, recognizing that *oral medications* refer to an act of administering medication rather than an instrument, and knowing too, the other ways in which medication can be administered, and that injection is one such way, enables the selection of the *injection* reading[42].

alilala chali he-slept-on-his-back	akihesabu counting	pau rafters/bread/clubs	na huku akifikiria while pondering	la kufanya what to do	(4.34)
dawa za kunywa oral medications	zinafanya kazi work	vizuri zaidi much better	kupita than	sindano needle/injection	(4.35)

Role of Bias For some of the results shown in table 4.23, ('*mkunga', 'kiongozi', 'kirukanjia'*) the best accuracy score obtained by the system is equal to the MFS baseline, and for those words where the result was better (and the sense bias was also correct), the MFS heuristic is implicitly considered during disambiguation. This first sense heuristic, where the correct sense is determined simply as that which is most frequent, is very important for supervised systems, with McCarthy, Koeling, Weeds & Carroll (2004) reporting that it frequently outperforms WSD systems even when they take the surrounding context into account, such as in the English all-words task in SENSEVAL2. This is due to the highly-skewed sense distribution common in natural language where one sense is much more frequent than the rest. In addition, Gale et al.'s (1992*b*) "one sense per discourse" observation means that only one sense of a word occurs for a given domain or discourse. Having informative priors about sense distribution is thus important for supervised systems in order to achieve performance better or at least equal to the

[42] The SWATWOL tag set has recently (at the time of writing this dissertation) been augmented to cover the following domains: Health, Physics, Chemistry and Language (linguistics).

MFS baseline. However, as Bruce (1995) states, parameters that affect disambiguation results are the test corpus, the target words and their degree of ambiguity, with Gale et al. (1992*a*) and Leacock et al. (1993) emphasizing that the outcome of a disambiguation experiment is more dependent on the target word rather than the disambiguation system itself. In light of these observations, there is a need for dynamic bias/prior determination for a word's senses that is specific for the domain and text type under test. This is particularly important in the absence of sense-tagged data which could be used as an approximation of the sense's true bias. It would also support the porting of WSD systems to different domains and corpora. As a solution to this problem, McCarthy, Koeling & Weeds (2004) have devised a system that ranks WordNet noun senses automatically by using thesauri created automatically from a raw corpus, coupled with WordNet-based similarity measures. With this system, they are able to determine the predominant sense for a given domain and text type as required. This provides reliable prior estimates which are useful for supervised WSD systems. However, as the results demonstrate, in the absence of a priori sense distribution, assuming no bias for any of the senses is a viable alternative, since the results obtained for both cases are comparable.

BBNs Most previous research using probabilistic models focuses on the simpler NB classifier. In this study, the merits of using a more powerful probabilistic model, namely the BBN is demonstrated. The BBNs, due to their intricate dependency modeling are better suited to natural language data which is characterised by high correlation in features, and this is supported by the better prediction results achieved using BBNs over NB classifiers. Also, unlike other WSD studies where BBNs have been used to model relationships between words to form a sort of semantic web or hierarchy (Wiebe et al. 1998, Ramakrishnan et al. 2004), the BBNs have been used in a more 'classical' setting where they express the dependencies and relationships inherent in commonly used feature sets. In addition, the prior knowledge supplied to the networks in terms of bias settings naturally allows them to default to the MFS, in the absence of additional information. This guarantees the best possible performance (baseline) even with minimal training data available.

4.5 Semantic Tagger for Swahili

For the WSD task, resolving 2-way ambiguity required computing the probability that the given test vector belongs to one of the two classes (representing each of the target noun's senses), and choosing the class that awards a higher probability as the right class, and hence the correct sense. By changing the classification configuration of the BBNs, the 15 classifiers can be used to achieve semantic tagging. In this case, rather than choose between two competing classes, all the classes are considered, and the one that gives the highest probability out of the 15, is chosen as the winning class. The test vector is then tagged with the corresponding class label. Formally, the semantic tagging task is described as follows:

$$c_{tag} = \arg\max_{c_i \in C} \Pr\left(d_t \mid c_i\right), \quad i = 1 \ldots 15$$

where C is the set of all 15 semantic classes, d_t is the test vector representing the target noun to be tagged, in context, and c_{tag} represents the winning class with whose label the target noun is tagged.

To test the proposed semantic tagging approach, 12 classes were chosen on the basis of availability of comparable training data sizes so as to have near-uniform priors and avoid biasing the result in favour of any class, and at the same time, allow the use of all available training data occurrences, during training. These classes are: HUMAN, LOCATION, TIME, INSTITUTION, ARTIFACT, FOOD, MONEY, SUBSTANCE, DRESS, VEHICLE, ANIMAL and UNIT. The training and test data that was used for test A (section 4.3.1) was reused for the tagging experiments, with the only difference being the change in the testing configuration as explained in the preceding paragraph. The results[43] obtained are presented in section 4.5.1.

[43] The results shown are based only on context window size +/- 2, without NP-chunking, as this was shown to achieve the best disambiguation results.

4.5.1 Results

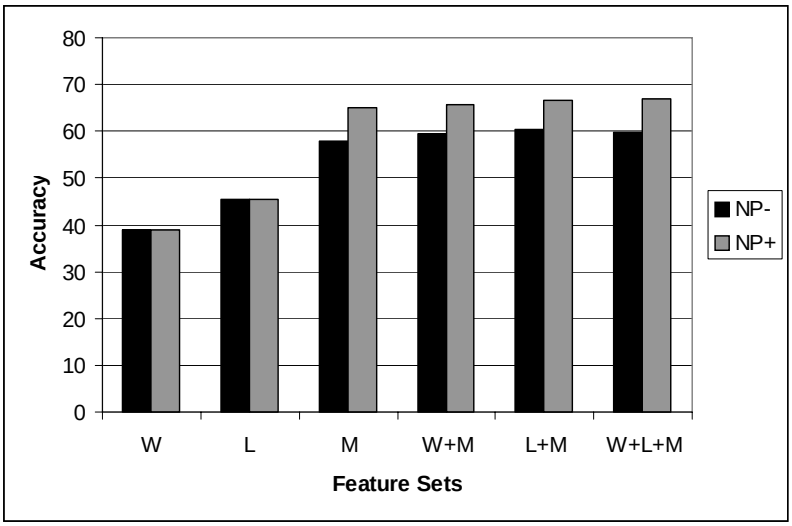

Figure 4.24: Accuracy results for semantic tagging

As shown in figure 4.24, two sets of experiments were done - one where the noun prefix (NP) feature was included (for feature set **M**) and the other where it was excluded. Since all senses of an ambiguous word have the same noun prefix, this feature was excluded in the WSD task as it was uninformative with respect to the sense. For tagging, this information may be important, and was included. However, the two sets of experiments were carried out with a view to assessing whether the Swahili noun prefix carries any semantic information that would be important for semantic classification, an issue that has generated much debate in Swahili linguistics as mentioned in section 3.3.1.

The results show that including the noun prefix feature yields an average increase in the accuracy of the tagger of 6.8 percentage points. While it would be impossible to make a conclusive statement regarding the role of the noun prefix with regard to semantic classification on the basis of this figure alone, it does provide empirical evidence which suggests that Swahili noun classes do contain a certain level of semantic coherence. This is especially so for class 1/2 which is largely HUMAN in composition, class 7/8 which is mainly comprised of ARTIFACTs and class 11 where most ABSTRACT nouns are found.

Like in all previous experiments, feature set **M** achieves the highest accuracy (64.9%), followed by **L** (45.3%) and lastly by **W** (39.0%). The highest overall accuracy, at 66.9% is achieved by a combination of all feature sets (**W+L+M**). The tagging accuracy is however much lower than that achieved for WSD for the same evaluation conditions, with the latter registering accuracies of 87.5% for **M**, 84.3% for **L** and finally 81.8% for **W**. The drop in performance is attributed to the increase in number of competing classifiers (classes) from 2 for WSD to 12 for tagging, without additional disambiguation information. With the tagging's 12-way ambiguity, the features' discriminatory power is considerably reduced, with some feature sets e.g. **W** not having sufficient discriminatory information to tell a majority of the classes apart. However, all feature sets significantly outperform the MFS baseline of 14%.

Despite the lower accuracy figures achieved for semantic tagging, this experiment has shown that it is possible to alter the test configuration of the BBN classifiers resulting in a semantic tagger. With improvement in the feature set to include global context as explained in section 4.4, the semantic tagging process can be used to provide default semantic tags for a Swahili lexicon. These could later be verified by hand. In addition, for new (unknown) words, the semantic tagger provides a better than chance heuristic in deciding the semantic properties for such words, and this could prove useful for other levels of linguistic processing such as morphological disambiguation and syntax parsing.

4.6 Summary

The evaluation of the Bayesian classifiers using standard WSD performance metrics has been presented in this chapter. The performance of the Bayesian classifiers surpasses that of the simple majority classifier, on all the standard performance metrics. The results obtained are thus satisfactory and promising, providing empirical justification of the WSD methodology employed in the study. The BBNs outperform the simple NB classifiers and this is attributed to their more sophisticated encoding of feature dependencies, unlike the independence assumption made by the latter. This characteristic is especially important

for natural language data, where features are highly correlated. The main research questions raised in section 3.6.1 have been addressed, with the conclusion that morphological and part-of-speech features collected from a small window of ± 2 is sufficient for Swahili WSD. However, with a dependency-type grammar for Swahili, it would be possible to gain valuable disambiguation information from verb-based feature sets. With regard to these verb-based feature sets, the study made a comparative analysis of the performance of classifiers trained using verbal semantic information obtained separately from WordNet and Levin's classes. It was shown that the latter provides more succinct disambiguation information for nouns and could be used either as an alternative to, or in conjunction with WordNet. The study also highlighted the need to edit the existing Swahili-English MRD in order to make it more usable for computational purposes. By altering the configuration of the BBN classifiers to include all of them in the classification of a test vector, a semantic tagger is obtained. The results obtained by this tagger are highly significant since despite the increase in sense granularity (from 2 to 12) and without a matching increasing in disambiguation information, the tagger's performance greatly exceeds the MFS baseline.

Chapter 5
Conclusion

The overall theme in this study is to advance the state of the art in LT for less-studied languages. This has been achieved by considering the problem of WSD, which is essential for language understanding applications, and which is considered to be one of the most challenging of all NLP research areas due to its reliance on a varied range of linguistic, statistical and real world knowledge.

The problem of WSD is addressed in the context of Swahili-English MT where it is viewed as that of choosing the right English translation for an ambiguous Swahili noun. The SOM algorithm is used in an exploratory phase to cluster occurrences of unambiguous nouns to obtain a semantic landscape of Swahili nouns. By using WordNet's noun classes as a semantic class building block, the automatically obtained semantic landscape is refined to yield fifteen major semantic classes, which are distinguishable on the basis of overtly-marked linguistic features for Swahili, and which form the building blocks for the WSD solution.

In total, the chosen methodology has been justified in terms of its theoretical foundations as well as the results obtained when the developed system is used to tag both ambiguous and unambiguous Swahili nouns with their appropriate semantic tags (senses) based on a given context. Given the simplicity of the feature set in use, the use of automatically-acquired training data and the reliance only on morphological analysis with minimal (surface) syntactic information, the results achieved are considered satisfactory and promising, since they surpass the simple majority classifier for both WSD and tagging. The results are especially promising for tagging, where accuracy increases from 14% to 66%, registering close to a five-fold increase over the majority classifier. This is especially significant given that tagging is the overall NLP goal of WSD.

5.1 Research contributions

The main contributions of this dissertation to LT research in general, and Swahili NLP in particular are:

- Creation of a word category map for Swahili nouns using the SOM algorithm. This map represents a semantic landscape for Swahili nouns that shows their distributional properties and semantic similarities given a set of text-based linguistic features. For each of the obtained categories, an analysis of the cluster properties shows what features are important for given categories. This information is very useful as it forms the foundation for subsequent semantic analysis for Swahili nouns. In this regard, the SOM has been used as a feature selector to determine the most powerful features for sense disambiguation.

- Automatic acquisition of annotated training data for WSD based on the obtained semantic category map. By identifying unambiguous member nouns for each of the semantic categories, occurrences of these nouns were extracted from the Swahili corpus and labelled with their class tag. This produced sufficient labelled data required for training the BBNs for WSD. In addition, the hand-tagged test corpus provides a gold standard for Swahili WSD that can be availed to the research community. This is an important contribution especially to the linguistic resources for Swahili, and will positively impact Swahili NLP capability.

- Comprehensive in-vivo testing for SWATWOL where the quality of its output is judged by the achieved WSD results. The achieved results vindicate the high accuracy reported for the morphological tagger and disambiguator. Some of the erroneous WSD results caused by wrong SWATWOL analyses provide useful feedback that can be used to further fine-tune SWATWOL's analysis and disambiguation engine. The improvement of WSD results with NP-chunking for those feature sets based on grammatical relations, offers empirical justification for the need to develop a dependency parser for Swahili.

- Development of a semantic tagger for Swahili nouns based on the SOM-induced semantic landscape. This tagger can be used to augment a Swahili lexicon with broad semantic tags that would support other levels of linguistic processing.

- Development of an unsupervised WSD system using BBNs. Due to its class-based approach, the system is able to disambiguate any noun whose senses are represented by different semantic classes, without having to build new word-specific discriminators for each additional noun, achieving general or broad-coverage WSD. Within a MT context, the WSD system can be incorporated as a TWS module.

- Design of a cross-lingual WSD methodology that does not make heavy demands on source language resource requirements, but instead exploits lexical resources available for other languages, specifically English (WordNet and Levin's verb classes), to provide vital semantic information for the source language. To this end, a computational semantic lexicon for Swahili verbs organised according to Levin's verb classes has been produced, and can be used to provide basic semantic categorization of Swahili verbs. In addition, the WSD system uses minimal computational resources i.e. morphological analyzer and disambiguator, without the need for full-fledged syntax parsing or bilingual corpora for TWS. This is a significant contribution especially for less-studied languages that have minimal computational and lexical resources, which is the case for most African languages. It demonstrates how to speed up LT research for these languages, by re-using existing resources for other languages, and concentrating only on critical source language analysis e.g. production of MRDs and alignment of these to existing computational lexicons such as WordNet, morphological analysis and corpora compilation.

5.2 Limitations

The main limitation of the developed solution is that the system relies heavily on the clusters or semantic categories obtained using the SOM algorithm, and this in turn determines the types of ambiguities the system can handle. In this case, the simplistic feature set allowed for the discovery of broad categories which represent coarse-grained ambiguities. Consequently, disambiguation is only possible if a word's senses belong to different categories. Disambiguation cannot be done for those words whose senses are of the same semantic type (see table 3.18). However, as determined in the analysis of inherent ambiguity types relevant for Swahili-English MT, most ambiguities are coarse-grained and the system may thus cover a significant proportion of ambiguous nouns for MT purposes. Nonetheless, highly polysemous nouns do occur within the language pair and would need to be disambiguated too.

5.3 Future work

The following areas present interesting research directions that if undertaken, would further improve the developed WSD solution:

- In the current configuration, the BBNs are trained in an unsupervised setting i.e. each classifier is trained only with positive examples for its class. Such a configuration is not optimized for classification since negative examples, which enhance a classifier's discriminatory power, are missing from the training data. The motivation for the current configuration was to gain an insight into the typical element of an individual class and to see what sorts of information are relevant in its definition independent of the other classes. This information is important when performing causal analysis to reveal what sorts of features are key in the definition of a particular class. It provides linguistic insight into the relationship between various linguistic features and semantics. In the proposed configuration, a single BBN, where the classification variable is contained within the network, would be trained using all the training data

for all classes. This would result in greater emphasis on the differences between the classes, and perhaps improve the classification performance of the system.

- The developed system uses only local context as a source of disambiguation information. Extending this to include global context - domain knowledge, topic and word associations (collocations and co-occurrences) is expected to improve results considerably. Related to this proposal, is the adoption of a two-tier approach to WSD where a word-specific classifier that takes advantage of context information that is specific and highly discriminative for a particular word is first employed in the disambiguation of the word. If the confidence threshold for this classifier is met, then its decision is taken to be the right one. However, if this is not the case, the system then falls back to the general class-based classifiers. In this way, the disambiguation algorithm attempts disambiguation by combining both types of disambiguation information - local and global. Further work could also entail using a variety of different feature sets to obtain the initial word category maps, as this may yield different classes which in turn has important consequences for the sense granularity and ambiguity coverage of the implemented WSD solution.

- In the absence of dependency parsing, simplistic modules were written to facilitate the acquisition of selectional preference information, by determining the direct objects and subjects of contextual verbs. The accuracy of this process is critical to the disambiguation performance of the WordNet and Levin-based feature sets. With a dependency parser for Swahili available, many errors in misclassification due to wrong processing of grammatical relations by the developed modules would be eliminated, and the true performance of these feature sets could be better determined. In addition, proper and complete editing of the existing Swahili-English dictionary and its alignment to WordNet and Levin's classes could have positive effects on the disambiguation performance, by ensuring that selectional preferences for more contexts are available for training, than is the case currently where this information is unavailable for approximately 12% and 26% of the verbs for WordNet and Levin respectively.

- Development of a named-entity recognition subsystem that would not only help in the disambiguation of proper vs. common nouns e.g. '*juma*', but would also provide useful features for the disambiguation of other classes. Identifying place names, food & beverage names, person names, disease names etc., would provide very reliable disambiguation cues for LOCATION, FOOD, PERSON and ABSTRACT classes, where such proper names occur within the context of an ambiguous noun. For example, in the fragment '*mto wa Tana*'(*river/pillow* of Tana), recognizing that '*Tana*' is the name of a river enables selection of the LOCATION sense (*river*) over the ARTIFACT sense (*pillow*).

- Given that the general methodology is applicable to any part of speech, the WSD system can be readily extended to cover other word categories, especially verbs. This would entail following the same procedure that was outlined for noun WSD. However, rather than use WordNet to supply nominal semantics, the developed semantic tagger could be used to provide this information. WordNet could then be used to supplement the tagger's information. However, due to the finer granularity of verb senses, careful selection of features would be required for the SOM clustering step, in order to obtain more and well-separated clusters, that would be sufficient to support resolution of the higher degree of ambiguity. In addition, since the methodology has been designed to be data-driven and thus language independent, it can be adapted to other less-covered languages, depending on their existing resources. At a minimum, a bilingual dictionary is required, and an additional requirement is that the second (target) language be necessarily one that has adequate resources, such as corpora and computational lexicons[44]. For those languages that have a reasonably sized monolingual corpus, the method can be applied directly as it was for Swahili. However, for those without such a corpus, the WSD method can be modified to take advantage of the target language corpus as a source of disambiguation information. This would entail matching the translated context of the ambiguous source language word to a

[44] The EuroWordNet project has the potential to increase the number and diversity of the linguistic and computational resources available to facilitate NLP of less-studied languages. [http://www.illc.uva.nl/EuroWordNet/]

target language corpus to identify the most probable target language sense, achieving disambiguation.

Bibliography

Agirre, E. & Martinez, D. (2000), Exploring automatic word sense disambiguation with decision lists and the web, *in* 'Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content'.

Agirre, E. & Martinez, D. (2001), Knowledge sources for word sense disambiguation, *in* R. M. Vaclav Matousek, Pavel Mautner & K. Tauser, eds, 'Proceedings of the Fourth International Conference on Text, Speech and Dialogue', Vol. 2166, Springer Verlag Lecture Notes in Computer Science series, Plzen (Pilsen), Czech Republic, pp. 1–10.

Aliferis, C. & Cooper, G. (1994), An evaluation of an algorithm for inductive learning of Bayesian Belief Networks using simulated data sets, *in* 'Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence', Seattle, WA, pp. 8–14.

Banfield, J. D. & Raftery, A. E. (1993), 'Model-based gaussian and non- gaussian clustering', *Biometrics* **49**, 803–821.

Bar-Hillel, Y. (1960), 'Automatic translation of languages', *Advances in Computers* .

Batibo, H. M. (1988), 'Root affixation rules in Zairean Kiswahili as evidence for earlier Bantu rules', *Journal of the Institute of Kiswahili Research* pp. 58–70.

Bernando, J. M. & Smith, A. F. M. (1994), *Bayesian Theory*, John Wiley.

Bick, E. (2002), *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press, Aarhus.

Boguraev, B. (1979), Automatic Resolution of Linguistic Ambiguities, PhD thesis, Computer Laboratory, Cambridge University, Cambridge, United Kingdom.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth and Brooks, Monterey, California.

Brill, E., Magerman, D., Marcus, M. & Santorini, B. (1990), Deducing linguistic structure from the statistics of large corpora, *in* 'Proceedings of the DARPA Speech and Natural Language Workshop', Morgan Kaufmann., San Mateo, CA, pp. 275–282.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R. & Roossin, P. (1990), 'A statistical approach to machine translation', *Computational Linguistics* **16**(2), 79–85.

Brown, P., Pietra, S. D., Pietra, V. D. & Mercer, R. (1991*a*), A statistical approach to sense disambiguation in machine translation, *in* 'Fourth DARPA Workshop on Speech and Natural Language', Pacific Grove, CA, pp. 146–151.

Brown, P., Pietra, S., Pietra, V. D. & Mercer, R. (1991*b*), Word sense disambiguation using statistical methods, *in* 'Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL-91)', Berkeley, CA, pp. 264–270.

Brown, P., Pietra, V. J. D., DeSouza, P. V., Lai, J. C. & Mercer, R. L. (1992), 'Class-based n-gram models of natural language', *Computational Linguistics* **18**(4), 467–479.

Bruce, R. (1995), A Statistical Method for Word Sense Disambiguation, PhD thesis, New Mexico state university, Las Cruces, New Mexico.

Buntine, W. (1991), Theory refinement on Bayesian networks, *in* B. D'Ambrosio, P. Smets & P. Bonissone, eds, 'Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence', Morgan Kaufmann Publishers, pp. 52–60.

Carlson, L. (1995), ValTer: Multilingual term bank system for terminology work, *in* 'Terminology in Advanced Microcomputer Applications : Proceedings of the 3rd TermNet Symposium : Recent Advances and User Reports, TermNet', Vienna, pp. 289–310.

Carlson, L. (1996), 'EAGLES Evaluation of Natural Language Processing systems. Evaluation of translators' aids.', Electronic References. Retrieved November, 2003.
URL: *http://www2.echo.lu.*

Chickering, D. M., Geiger, D. & Heckerman, D. (1994), Learning Bayesian networks is NP-hard, Technical Report MSR-TR-94-17, Microsoft Research, Microsoft Corporation.

Choueka, Y. & Lusignan, S. (1985), 'Disambiguation by short contexts', *Computers and the Humanities* **19**, 147–158.

Ciaramita, M., Hofmann, T. & Johnson, M. (2003), Hierarchical semantic classification: Word sense disambiguation with world knowledge, *in* 'Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)', Acapulco, Mexico.

Contini-Morava, E. (1997), Noun classification in Swahili: A cognitive-semantic analysis using a computer database, *in* R. K. Herbert, ed., 'African Linguistics at the Crossroads: Papers from Kwaluseni, 1st World Congress of African Linguistics', Cologne: Rüdiger Köppe, Swaziland, pp. 599–628.

Cooper, G. & Herskovits, E. (1992), 'A Bayesian method for the induction of probabilistic networks from data', *Machine learning* **9**, 309–347.

Cottrell, G. (1985), A Connectionist Approach to Word Sense Disambiguation, PhD thesis, Department of Computer Science, University of Rochester, Rochester, N.Y, USA.

Cowie, J., Guthrie, J. & Guthrie, L. (1992), Lexical disambiguation using simulated annealing, *in* 'Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)', Vol. 1, Nantes, France, pp. 359–365.

Cruse, A. D. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge, England.

Dagan, I. & Itai, A. (1994), 'Word sense disambiguation using a second language monolingual corpus', *Computational Linguistics* **20**(4), 563–596.

Dawid, A. P. (1984), 'Present position and potential developments: Some personal views, statistical theory, the prequential approach', *Journal of the Royal Statistical Society* **A 147**, 178–292.

Diab, M. & Resnik, P. (2002), An unsupervised method for word sense tagging using parallel corpora, *in* 'Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)', Philadelphia, PA, pp. 255–262.

Eizirik, L., Barbosa, V. & Mendes, S. (1993), 'A Bayesian-network approach to lexical disambiguation', *Cognitive Science* **17**, 257–283.

Fellbaum, C. (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge (MA).

Fraley, C. & Raftery, A. (2002), 'Model-based clustering, discriminant analysis and density estimation', *Journal of the American Statistical Association* **97**, 611–631.

Gale, W., Church, K. & Yarowsky, D. (1992*a*), Estimating upper and lower bounds on the performance of word sense disambiguation programs, *in* 'Proceedings of the 30th Conference of the Association for Computational Linguistics', Newark, Delaware, pp. 249–256.

Gale, W., Church, K. & Yarowsky, D. (1992*b*), One sense per discourse, *in* 'In Proceedings of the DARPA Speech and Natural Language Workshop', Harriman, NY, pp. 233–237.

Gale, W., Church, K. & Yarowsky, D. (1992*c*), Using bilingual materials to develop word sense disambiguation methods, *in* 'Proceedings of the Fourth Inter-National Confer-

ence on Theoretical and Methodological Issues in Machine Translation', Montreal, Canada, pp. 101–112.

Gonzalo, J., Chugur, I. & Verdejo, F. (2002), Polysemy and sense proximity in the senseval-2 test suite, *in* 'Proceedings of Word Sense Diasmbiguation: Recent Successes and Future Directions', University of Pennsylvania, Pennsylvania,.

Grefenstette, G. (1994), *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, Boston, MA.

Guthrie, J., Guthrie, L., Wilks, Y. & Aidinejad, H. (1991), Subject dependent co-occurrence and word sense disambiguation, *in* 'Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics', Berkeley, California, pp. 146–152.

Hayes, P. (1977), Some association-based techniques for lexical disambiguation by machine, Technical Report 25, Department of Computer Science, University of Rochester, N.Y, USA.

Heckerman, D. (1996), A tutorial on learning with Bayesian networks, Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, Microsoft Corporation.

Heckerman, D., Geiger, D. & Chickering, D. M. (1995), 'Learning Bayesian networks: The combination of knowledge and statistical data', *Machine Learning* **20**(3), 197–243.

Heckerman, D., Mamdani, A. & Wellman, M. (1995), 'Real-world applications of Bayesian networks', *Communications of the ACM* **38**(3).

Hirst, G. (1987), *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, Cambridge, UK.

Honkela, T. (1997), Self-Organizing Maps in Natural Language Processing, PhD thesis, Department of Computer Science and Engineering, Helsinki University of Technology.

Honkela, T., Kaski, S., Lagus, K. & Kohonen, T. (1997), WEBSOM - Self-Organizing Maps of document collections, *in* 'Proceedings of WSOM'97, Workshop on Self-Organizing Maps', Helsinki University of Technology, Finland, pp. 310–315.

Howard, R. A. & Matheson, J. E., eds (1984), *READINGS on the Principles and Applications of Decision Analysis*, Strategic Decisions Group, Menlo Park, California.

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 the 16th International Conference on Computational Linguistics', Copenhagen, pp. 568–573.

Hurskainen, A. (1999), 'SALAMA: Swahili Language Manager', *Nordic Journal of African Studies* **8**(2), 139–157.

Hurskainen, A. (2004*a*), Computational testing of five Swahili dictionaries, *in* 'Proceedings of the 20th Scandinavian Conference of Linguistics', Helsinki, Finland.

Hurskainen, A. (2004*b*), Optimizing disambiguation in Swahili, *in* 'Proceedings of COLING-04, the 20th International Conference on Computational Linguistics', Geneva, Switzerland, pp. 254–260.

Hutchins, W. J. (1995), Machine translation: A brief history, *in* E. F. K. Koerner & R. E. Asher, eds, 'Concise History of the Language Sciences : From the Sumerians to the Cognitivists', Pergamon, New York.

Ide, N. (2000), 'Cross-lingual sense determination: Can it work?', *Computers and the Humanities: Special Issue on the Proceedings of the SIGLEX/SENSEVAL Workshop* **34**(1-2), 223–234.

Ide, N. & Véronis, J. (1998), 'Introduction to the special issue on word sense disambiguation: The state of the art', *Computational Linguistics* **24**(1), 1–40.

Jensen, F. (1996), *An Introduction to Bayesian Networks*, UCL Press, London.

Jurafsky, D., Martin, J. H. & Kehler, A. (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Upper Saddle River (N.J.): Prentice Hall.

Kaplan, A. (1955), 'An experimental study of ambiguity and context', *Mechanical Translation* **2**(2), 39–46.

Karlsson, F. (1990), Constraint grammar as a framework for parsing running text., *in* H. Karlgren, ed., 'COLING-90. Papers Presented to the 13th International Conference on Computational Linguistics', Vol. 3, Helsinki, pp. 168–173.

Karlsson, F., Voutilainen, A., Heikkilä, J. & Antilla, A. (1995), *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, Berlin.

Kelly, E. & Stone, P. (1975), 'Computer recognition of English word senses', *North Holland Linguistics Series* **3**.

Kikui, G. (1999), Resolving translation ambiguity using non-parallel bi-lingual corpora, *in* 'Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language Processing', College Park, Maryland.

Kilgarriff, A. (1997), 'I don't believe in word senses', *Computers and the Humanities* **31**(2), 91–113.

Kilgarriff, A. (1998), SENSEVAL: An exercise in evaluating word sense disambiguation programs, *in* 'Proceedings of the First International Conference on Language Resources and Evaluation', Granada, Spain, pp. 581–585.

Kohonen, T. (1995), *Self-Organizing Maps*, Springer-Verlag, Heidelberg, Berlin.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H. & Grünwald, P. (2000), 'On predictive distributions and Bayesian networks', *Statistics and Computing* **10**, 39–54.

Koskenniemi, K. (1983), Two-level morphology: A general computational model for word-form recognition and production, Technical Report 11, Department of General Linguistics, University of Helsinki.

Koutsoudas, A. & Korfhage, R. (1956), 'MT and the problem of multiple meaning', *Mechanical Translation* **2**(2), 46–51.

Krause, P. (1998), 'Learning probabilistic networks', *Knowledge Engineering Review* **13**, 321 – 351.

Lagus, K. & Airola, A. (2001), Analysis of functional similarities of Finnish verbs using the Self-Organizing Map, *in* 'ESSLLI'01 Workshop on The Acquisition and Representation of Word Meaning', Helsinki, Finland.

Leacock, C., Chodorow, M. & Miller, G. A. (1998), 'Using corpus statistics and WordNet relations for sense identification', *Computational Linguistics* **24**(1), 147–166.

Leacock, C., Towell, G. & Voorhees, E. (1993), Corpus-based statistical sense resolution, *in* 'Proceedings of the ARPA Workshop on Human Language Technology', Morgan Kaufman, San Francisco, CA.

Lenci, A. (2001), Building an ontology for the lexicon: Semantic types and word meaning, *in* P. A. Jensen & P. Skadhauge, eds, 'Proceedings of the First International OntoQuery Workshop', Kolding, pp. 43–56.

Lesk, M. E. (1986), Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, *in* 'Proceedings of the SIGDOC Conference', Toronto, Ontario, Canada, pp. 24–26.

Levin, B. (1993), *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press.

Liddy, E. (1998), Knowledge discovery using KNOW-IT, *in* 'Proceedings of the 1998 IEEE Information Technology Conference'.

Lindén, K. & Lagus, K. (2002), Word sense disambiguation in document space, *in* 'Proceedings of IEEE International Conference on Systems, Man and Cybernetics', Hammamet, Tunisia.

Luk, A. (1995), Statistical sense disambiguation with relatively small corpora using dictionary definitions, *in* 'Proceedings of the 33rd Meeting of the Association for Computational Linguistics (ACL-95)', Cambridge, MA, pp. 181–188.

Lyons, J. (1977), *Semantics*, Cambridge University Press, Cambridge, England.

Madigan, D., Raftery, A., Volinsky, C. & Hoeting, J. (1996), Bayesian model averaging, *in* 'Proceedings of the AAAI Workshop on Integrating Multiple Learned Models', Portland, OR.

Manning, C. & Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT press, Cambridge, MA.

Màrquez, L. (2000), Machine Learning and Natural Language Processing, Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Informàtics (LSI), Universitat Politecnica de Catalunya (UPC), Barcelona, Spain.

Masterman, M. (1957), 'The thesaurus in syntax and semantics', *Mechanical Translation* **4**, 1–2.

McCarthy, D., Koeling, R. & Weeds, J. (2004), Ranking WordNet senses automatically, Technical Report CSRP 569, Department of Informatics, University of Sussex.

McCarthy, D., Koeling, R., Weeds, J. & Carroll, J. (2004), Finding predominant senses in untagged text, *in* 'Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics', Barcelona, Spain.

McCord, M. C. (1990), Slot grammar: A system for simpler construction of practical natural language grammars, *in* R. Studer, ed., 'Natural Language and Logic: Proceedings of the International Scientific Symposium', Springer, Berlin, Heidelberg, pp. 118–145.

McDonald, J., Plate, T. & Schvaneveldt, R. (1990), Using pathfinder to extract semantic information from text, *in* R. W. Schvaneveldt, ed., 'Pathfinder Associative Networks: Studies in Knowledge Organization', Ablex, Norwood, N.J, USA.

McRoy, S. W. (1992), 'Using multiple knowledge sources for word sense discrimination', *Computational Linguistics* **18**(1), 1–30.

Michie, D., Spiegelhalter, D. & Taylor, C. (1994), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York.

Mitchell, T. (1997), *Machine Learning*, McGraw-Hill, New York.

Mooney, R. J. (1996), Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning, *in* E. Brill & K. Church, eds, 'Proceedings of the Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Somerset, New Jersey, pp. 82–91.

Mulokozi, M. M. (2002), 'Kiswahili as a national and international language'. Retrieved August, 2002.
URL: *www.helsinki.fi/hum/aakkl/documents/kiswahili.pdf*

Myllymäki, P., Silander, T., Tirri, H. & Uronen, P. (2002), 'B-course: A web-based tool for Bayesian and causal data analysis', *International Journal on Artificial Intelligence Tools* **11**(3), 369–387.

Nadas, A. (1983), 'A decision-theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood', *IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP)* **31**(4), 814–817.

Ng, H. & Lee, H. (1996), Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach, *in* 'Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL-96)', Santa Cruz, CA., pp. 40–47.

Ng, H. T. (1997), Exemplar-based word sense disambiguation: Some recent improvements, *in* 'Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing'.

Ng'ang'a, W. (2003*a*), Automatic word sense disambiguation - Kiswahili nouns, *in* 'Proceedings of the 4th World Congress of African Linguistics', New Brunswick, NJ, USA.

Ng'ang'a, W. (2003*b*), 'Semantic analysis of Kiswahili words using the Self-Organizing Map', *Nordic Journal of African studies* **12**(3), 405–423.

Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Pedersen, T. & Bruce, R. (1997), Distinguishing word senses in untagged text, *in* 'Proceedings of the Second Conference on Empirical Methods in Natural Language Processing', Providence, RI, pp. 197–207.

Pustejovsky, J. (1995), *The Generative Lexicon*, MIT Press.

Quillian, R. (1961), A design for an understanding machine, *in* 'Colloquium on Semantic Problems in Natural Language', King's College, Cambridge University, Cambridge, United Kingdom.

Rabiner, L. & Juang, B. (1993), *Fundamentals of Speech Recognition*, Prentice Hall, Upper Saddle River, NJ.

Ramakrishnan, G., Prithviraj, B. P., Deepa, A., Bhattacharya, P. & Chakrabarti, S. (2004), Soft word sense disambiguation, *in* 'Proceedings of the Second International WordNet Conference', Brno, Czeck Republic, pp. 291–298.

Resnik, P. (1999), Disambiguating noun groupings with respect to WordNet senses, *in* S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann & D. Yarowsky, eds, 'Natural Language Processing Using Very Large Corpora', Kluwer Academic Publishers, Boston, M.A, pp. 77–98.

Resnik, P. & Yarowsky, D. (1997), A perspective on word sense disambiguation methods and their evaluation, *in* 'ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?', Washington, D.C., USA, pp. 79–86.

Resnik, P. & Yarowsky, D. (1999), 'Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation', *Natural Language Engineering* **5**(2), 113–134.

Rigau, G., Atserias, J. & Agirre, E. (1997), Combining unsupervised lexical knowledge methods for word sense disambiguation, *in* '35th Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics (ACL/EACL-97)', Madrid, Spain, pp. 48–55.

Rodd, J., Gaskell, M. & Marslen-Wilson, W. D. (2000), The advantages and disadvantages of semantic ambiguity, *in* L. R. Gleitman & A. K. Joshi, eds, 'Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society', Lawrence Erlbaum Associates, Mahwah, New Jersey, pp. 405–410.

Ruohotie, R., Nokelainen, P., Tirri, H. & Silander, T. (2001), *Modeling Individual and Organizational Prerequisites of Professional Growth - Papers Presented at International Conferences 1999-2001*, Hämeenlinna Häme Polytechnic.

Schütze, H. (1993), Word space, *in* J. Hanson, J. Cowan & C. Giles, eds, 'Advances in Neural Information Processing Systems', Vol. 5, Morgan Kaufmann, San Mateo, CA, pp. 895–902.

Schütze, H. (1998), 'Automatic word sense discrimination', *Computational Linguistics* **24**(1), 97–124.

Sewangi, S. S. (2001), Computer-Assisted Extraction of Terms in Specific Domains: The Case of Swahili, PhD thesis, University of Helsinki.

Spiegelhalter, D., Dawid, A. P., Lauritzen, S. & Cowell, R. (1993), 'Bayesian analysis in expert systems', *Statistical science* **8**, 219–282.

Tapanainen, P. (1996), The constraint grammar parser CG-2, Technical Report 27, Department of General Linguistics, University of Helsinki.

Véronis, J. & Ide, N. (1990), Word sense disambiguation with very large neural networks extracted from machine readable dictionaries, *in* 'Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)', Vol. 2, Helsinki, Finland, pp. 389–394.

Vesanto, J. & Alhoniemi, E. (2000), 'Clustering of the Self-Organizing Map', *IEEE Transactions on Neural Networks* **11**(3), 586–600.

Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. (2000), Self-Organizing Map in Matlab: The SOM toolbox, *in* 'Proceedings of the Matlab DSP Conference, Espoo, Finland'.

Waltz, D. & Pollack, J. (1985), 'Massively parallel parsing: A strongly interactive model of natural language interpretation', *Cognitive Science* **9**, 51–74.

Weaver, W. (1949), Translation, *in* 'Machine Translation of Languages, 1955', MIT Press, Cambridge, MA.

Weiss, S. M. & Kulikowski, C. A. (1991), *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*, Morgan Kaufmann, San Mateo (CA).

Wiebe, J., O'Hara, T. & Bruce, R. (1998), Constructing Bayesian networks from WordNet for word sense disambiguation: Representation and processing issues, *in* 'Proceedings of COLING-ACL '98 Workshop on the Usage of WordNet in Natural Language Processing Systems, Association for Computational Linguistics', Montreal, Canada.

Wilks, Y. (1975), 'A preferential, pattern-seeking semantics for natural language inference', *Artificial Intelligence* **6**, 53–74.

Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T. & Slator, B. (1990), Providing machine tractable dictionary tools, *in* J. Pustejovsky, ed., 'Semantics and the Lexicon', MIT Press, Cambridge, M.A, USA.

Yarowsky, D. (1992), Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, *in* 'Proceedings of the Fourteenth International Conference on Computational Linguistics', Nantes, France, pp. 454–460.

Yarowsky, D. (1993), One sense per collocation, *in* 'Proceedings of ARPA Human Language Technology Workshop', Princeton, New Jersey, pp. 266–271.

Yarowsky, D. (1995), Unsupervised word sense disambiguation rivaling supervised methods, *in* 'Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics', Cambridge, M.A, pp. 189–196.