

On Efficient Creation of Machine Translation Systems

K. Narayana Murthy

Department of Computer and Information Sciences,

University of Hyderabad,

Hyderabad, 500 046, INDIA

email: knmcs@uohyd.ernet.in

Abstract

The information revolution and the Internet have made vast amounts of useful information freely available, crossing the barriers of space and time. Yet, information is not equally available for all, as oftentimes information is not available in a language that people know. This raises many issues and possibilities. We can as well create content in many different languages. We can create multi-lingual information sources. We can also use translation tools and make the information stored in one language available for people who know a different language, thus paving the way for multi-lingual information retrieval. This paper discusses some of these issues with specific emphasis on efficient creation of automatic translation tools.

It has been well recognized that general purpose, fully automatic, high quality translation is extremely difficult to achieve in practice. In many circumstances, however, accessing information in one's own language is more important than producing perfect translations. Also, computers have become ever faster and capable of storing and processing more and more data. Statistical techniques based on large corpora have become viable. Multi-lingual information retrieval is becoming attractive and both the demand and opportunities for machine translation are on the increase.

After a brief sketch of the current scenario and the problems and issues, we focus on developing efficient machine translation systems. Speed and efficiency are especially important in embedded systems where machine translation would be an embedded component. We develop a hybrid architecture for machine translation and show how a judicious combination of available linguistic knowledge and resources on the one hand, and corpus based statistical techniques on the other, can lead to very efficient machine translation systems. The proposed architecture is modular and the various modules are amenable for more or less independent and parallel development. We also discuss how corpus requirements can be reduced and how bootstrapping can be used for faster development of translation systems.

1 Introduction

The information revolution and the Internet have made vast amounts of useful information freely available, crossing the barriers of space and time. Yet, information is not equally available for all, as oftentimes information is not available in a language that people know. This raises many issues and possibilities. One possibility would be to create the same content in many different languages. Content could be created independently in each language. Alternatively, content already created in one language can be translated into other languages, either manually or through automatic translation tools. It is also conceivable to create multi-lingual content - where each unit of information within a document is available in many different languages as opposed to different versions of the same document for different languages. In all these cases, there is a kind of redundancy of information and an increased burden in terms of content creation as well as storage. Another better possibility, therefore, could be to create and store content in one or a few languages and develop translation tools for accessing this content in other languages. If translation is resorted to at both the input and output sides, complete transparency can be achieved. For example, one could send a query in a particular language and get the response also in the same language, while the information itself was available in a different language altogether. In any case, both the demand and opportunities for automatic translation are increasing.

High quality automatic translation has not been easy to achieve in practice in general. Machines have neither common sense nor world knowledge. They cannot understand language the same way we humans can. Despite all the developments in the fields of Artificial Intelligence and Machine Learning, computers today fall far short of human like translation capabilities. In some restricted applications, quality requirements can be relaxed to an extent. For example, in a key word based information retrieval system, it is enough if the keywords are properly translated. However, minimum quality requirements have to be enforced in most practical systems and the only way out would be to involve the human being at some stage or the other. While pre-editing is rarely acceptable, post-editing can be resorted to if quality of output has to be maintained. Interactive translation is yet another possibility - the machine seeks human help during translation as and when it is unable to make the right decisions. With the increasing use of large corpora and statistical methodologies, it is becoming increasingly possible to make educated guesses and go ahead without human intervention.

There are a large number of languages in the world and not all are in an equally advanced state of development in terms of technology. As an example, the 1991 census of India lists 18 Scheduled languages with constitutional recognition and 96 non-scheduled languages, the last one of which is 'the total of all other languages spoken by less than 10,000 people across India'. Thus even the names of these languages are not known! There are many languages where we do not even have dictionaries or grammars. Language diversity, language endangerment and language planning are large and complex subjects that need very careful study. Preservation of endangered languages and culture, development of minor languages, development of technology for major languages, may all

have to go on simultaneously.

There is an urgent need for developing acceptable standards at the character and glyph level encodings for various languages of the world to ensure inter-operability and inter-translatability. Absence of glyph standards have made character encoding standards meaningless, as only font encodings can be rendered by tools such as editors and browsers. Browsers understand only bytes, not whole characters and thus often render Indian languages wrongly.

In the rest of this paper, we focus on machine translation systems. We focus both on the efficiency of translation and the efficiency or speed with which such systems can be developed.

2 Automatic Translation

Translation is a meaning preserving transformation from one language (called the source language - SL) to another language (called the target language - TL). In other words, while the lexical items and their arrangement in terms of the structure of sentences may vary from SL to TL text, the meaning of the text should be preserved. When people translate, they usually read and ‘understand’ the SL text first and only then try to create a TL text accordingly. Thus one possible way of modeling translation could be as a three phase process - analysis of the meaning of the source language text, transfer of lexical items from SL to TL, and generation of the TL text keeping the syntax of the TL in mind. Dividing the process of translation into three distinct phases in this manner is surely a simplification, but it is useful for our treatment of translation here.

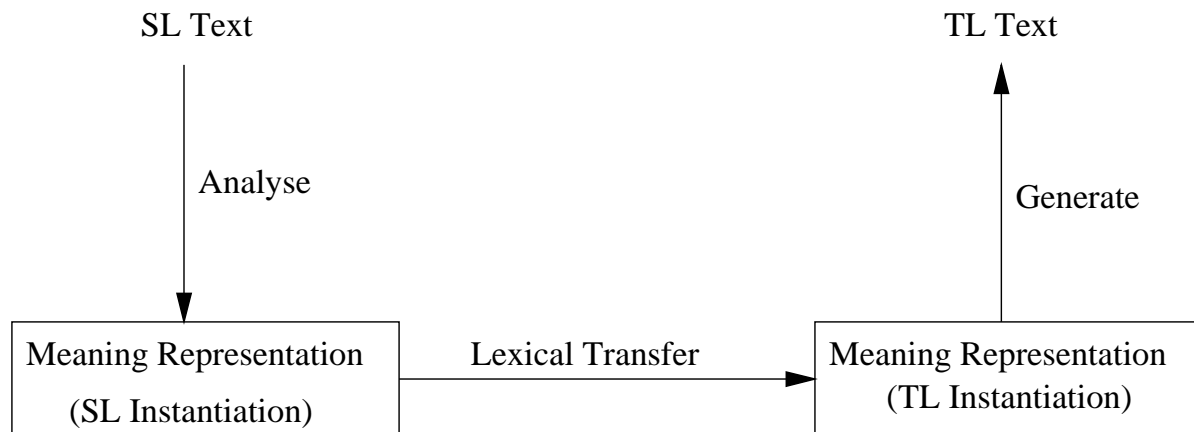


Figure 1. Machine Translation

Target text that depicts a changed or distorted meaning can hardly be considered as ‘valid’ translation. Thus, perhaps the most important aspect of translation is meaning preservation. This presupposes that the meaning of the source text can be analysed and understood. This is a natural expectation from human translators but when it comes to automatic or machine translation (MT), things are very different. Neither classical theories of meaning nor theories of formal semantics have developed to a stage where they can be fruitfully applied for machine translation.

‘Meaning’ is a vast and deep subject about which we do not know enough as yet. There are certainly different levels of meaning. Classical theories of meaning include 1) the Referential Theory, which says that the meaning of an expression is the actual object it refers to, 2) the mentalistic Theory, which says that the meaning of an expression is the idea associated with the expression in the mind of the speaker, and 3) the Use Theory of meaning, which asserts that the meaning of an expression is determined by its use in the language community. Indian philosophers have recognized three distinct levels of meaning - ‘abhida’ (primary, literal or dictionary meaning), ‘lakSaNa’ (figurative or metaphorical meaning), and ‘vyanjana’ (emotive, associative or suggestive meaning) [1]. Further, four conditions have been stipulated for understanding the meaning of an expression: 1) ‘aakaMkSa’ or expectation (ex. a transitive verb expects an object), 2) ‘yogyata’ or compatibility (ex. ‘banana’ can be an object of ‘eat’), 3) ‘sannidhi’ or co-occurrence (ex. In ‘follow the thief’ the expression ‘the thief’ fixes one meaning for the word ‘follow’ as distinguished from, say, ‘follow Mahatma Gandhi’s principles’ or ‘follow my lecture’), and, 4) ‘taatparya jnaana’ or context and speaker’s intentions (ex. a teacher in a classroom says ‘Are you following?’). Thus meaning merges seamlessly into the realm of intentions, goals, plans, strategies and other aspects of human cognition.

Depending on the depth of analysis and understanding, the translator may choose anything from simple and more or less direct transfer to ‘transcreation’ where the TL text can be legitimately viewed as new creation in itself. There can be significant variations in the degree of originality and creativity in translation. Apart from preserving meaning, the translator may choose to preserve style and effect too.

In depth and human-like understanding by machine has been shown to be possible under certain conditions [2]. Nevertheless, in depth understanding of natural language in realistic and general circumstances has remained a distant dream and no practical MT system can hope to achieve human-like understanding of the meaning of the source language text. One approach to MT is the interlingua approach, where translation from SL to TL is done through an intermediate, universal representation language called interlingua. Finding a good intermediate and universal representation is hard. At the other extreme, attempts to develop MT systems by more or less direct lexical transfer without any significant analysis of the SL text have not given very good results either. Another possibility is to focus more on lexical and syntactic structure. Thus a practical view of MT is simply to identify the right TL words and place in the right order. We shall argue here that structure and meaning are not entirely divorced from one another and, viewed appropriately, structure depicts meaning.

3 Structure as Meaning

‘Form follows function’ and syntax should really be viewed as a rendering of meaning at some appropriate level of linguistic description. After all, even the basic grammatical categories are really nothing more than ‘general meanings’. Thus nouns are things, verbs are actions or state descriptions, adjectives are properties or attributes of things and so on. It is unfortunate that the mainstream research in generative syntax has not given meaning and its relation to structure the treatment they deserve. Syntax is not viewed as semantics at a very high level of generalization, as it should ideally be. See [3] for an excellent critique of the Chomskian school of linguistics.

Classical grammars have divided syntax into three aspects - order, concord (or agreement) and government (or dependency). Thus, for a sentence to be grammatical, word order, agreement as well as the thematic roles and their dependency relations must all be valid. If we can analyse these aspects of structure in a given sentence, we would be able to specify who did what to whom, where, when, why, etc. in the given sentence. In this sense, we would have obtained the basic, literal and direct meaning of the sentence. Thus structure is nothing but meaning at an appropriate level of linguistic description. Since automating syntax is much more practicable than automating full semantics, it would be possible to build machine translation systems that analyse the structure of SL sentences, treat this as basic, literal meaning, and transfer this meaning to the TL. Deeper aspects of semantic correctness and appropriateness and style will not be automated but will simply be left for a human post-editor to work on manually. This is the view that we shall take in this paper.

Theories of syntax typically deal with the syntax of individual sentences, the structure of discourse being again a major area where we are nowhere close to practical application. We assume that sentences exist as identifiable and legitimate units of text in all the languages of our concern and work with one sentence at a time. This is what current circumstances permit as well as demand.

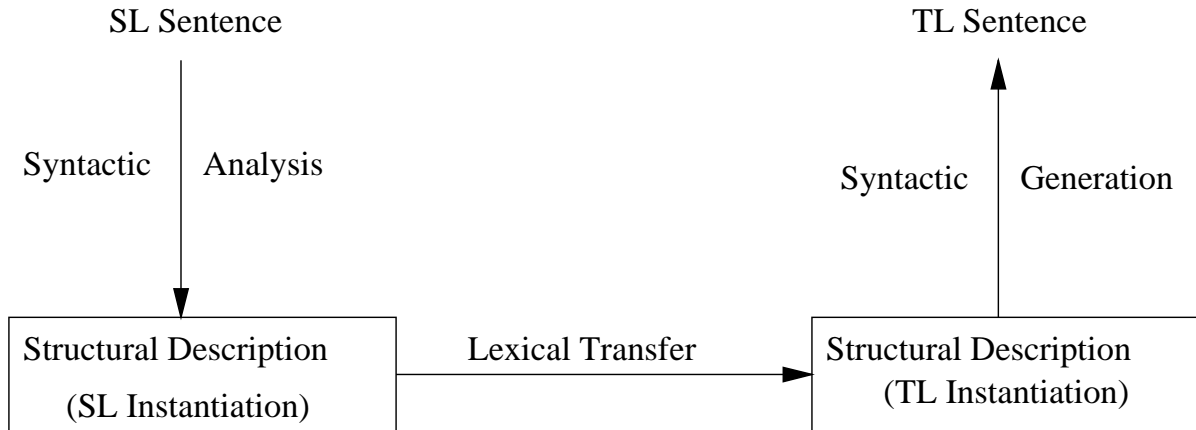


Figure 2. MT: Structure as Meaning

A syntactic analyser takes one sentence as input and produces as output a description of the structure of the given sentence. What kind of a description of structure should we look for? If we wish to view structural descriptions as basic meaning, tree structures will not be enough. Trees are good for depicting linear (word order) and hierarchical (constituent structure) relationships but are inherently not well suited for describing functional or thematic dependencies amongst the various constituents in a sentence. It is the dependency structure that tells us who did what to whom etc. and therefore the primary link to meaning. These functional roles have been called *kaarak* roles in the PaaNini’s treatment of the grammar of Sanskrit. These roles are semantic in nature and largely language independent. Universal Clause Structure Grammar (UCSG) [4] is a computational framework for analysing these structures from natural languages. UCSG has been designed to work well for positional languages such as English as well as for relatively free word order languages. The F structure of UCSG assigns thematic roles to all the complements as well as adjuncts in each clause in a sentence and also shows the inter-dependencies between the various clauses in a sentence.

Constituents of a sentence that get assigned to specific functional roles could be isolated words but often they are whole groups of words. We shall call these typically contiguous groups of words in a sentence as ‘word groups’. Word groups behave at atomic units in functional structure and take on one functional role each. We could have called them ‘phrases’ but we prefer the former terminology to avoid confusion. Linguists have come to use the term ‘phrase’ in very different senses. Thus when linguists talk of an np and a vp in a sentence as phrases, the vp may include not only the verb but also complements and adjuncts. On the other hand, a ‘verb group’ would only denote the sequence of auxiliary verbs and the main verb as in ‘has been writing’.

Since word groups behave as atomic units at functional structure level, syntactic anal-

ysis can be made very efficient if we build these word groups right in the beginning. At all subsequent stages of parsing, a ‘sentence’ could be viewed as a sequence of word groups rather than as a sequence of words. This reduces the effective length of the sentence and thus makes parsing much more efficient.

Further, it is extremely important to make a very clear distinction between phrases and clauses. Clauses are at the same level of linguistic description as simple sentences whereas word groups are simply groups of words that have a unitary meaning. Clauses include verbs and the associated thematic roles and thus exhibit dependency relationships. Phrases or word groups, on the other hand, are inherently much simpler. It would not be economical, either in linguistic terms or in computational terms, to mix up the two levels of description as most current theories do.

A noun group, for example, typically involves a specifier, zero, one or more adjectival modifiers and a head noun. The order of words in a word group is usually significant - changing the word order changes the meaning or renders the group anomalous. This is true of both positional languages such as English and the so called relatively free word order languages. Even in free word order languages, there is no real freedom of word order within word groups. In fact the freedom of order applies only for the entire word groups themselves. Thus linear precedence is an essential characteristic of phrase level analysis. What about hierarchical relationships? Tree structure tend to suggest that there is indeed a hierarchical structure in word groups. We view hierarchy in more semantic terms and consider hierarchical relationships as indicative of part-whole or general-specific relationships. For example, if we can identify the modifier-modified relationships in a noun group such as ‘water meter cover adjustment screw’, we could consider that to be a kind of hierarchical structure. Current theories of syntax fall far short of helping us to identify modifier-modified or such other semantic relationships and we do not see much point in drawing elaborate tree structures where there is no hierarchical structure that can actually be captured. In our model, we do not even attempt to produce a detailed internal structure of word groups. We instead contend with mere identification of word groups and subsequent assignment of appropriate functional roles. This a kind of ‘doing more by doing less’ strategy. By avoiding hierarchical nesting and limiting ourselves to mere recognition of word groups, we can parse efficiently with only Finite State Machine power.

Another well known principle that is rarely exploited in linguistic theories is the principle of locality. The functional structure of a clause is essentially local to that clause. The participating word groups of a clause do not normally cross the clause boundaries. There are well defined conditions for sharing, displacement or ellipsis of functional roles and except insofar as these rules permit, no group can trespass the boundaries of its clause. The significance of the locality principle is that functional structure of a sentence can be analysed clause by clause. If a sentence has ‘c’ clause each of which exhibits ‘r’ functional roles, the locality principle reduces the problem of assigning ‘c * r’ roles to ‘c * r’ word groups to the problem of ‘c’ times assigning ‘r’ roles to ‘r’ groups. Since the assignment problem is exponential in nature, factoring out ‘c’ leads to a very significant reduction in problem complexity for multi-clause problems.

While the locality principle itself is well known, linguistic models have not exploited the principle, perhaps because of the perceived difficulty in identifying clause boundaries. UCSG has demonstrated that clause boundaries can be identified and very efficiently [4]. In UCSG noun groups are ignored and only verb groups and sentinels are used for clause structure analysis. Thus the effective length of the ‘sentence’ used will be very small, about one third of the length of the original sentence on the average for English. Also, we need no more powerful a grammar system than Context Free Grammar. A very small number of grammar rules suffice and the overall algorithm is very simple and efficient. Having analysed the clause structure, UCSG shows that by working from whole to part instead of working from left to right, we can take care of all the inter-clause dependencies and carry out functional structure analysis essentially one clause at a time.

In summary, we propose a parser based translation system. The parser works one sentence at a time and first obtains all potential word groups in the given sentence. Then the clause structure of the sentence is analysed. Finally, functional roles are assigned to the various word groups in each clause. The structure so obtained is treated as a representation of basic meaning and the meaning is transferred to the target language through lexical transfer and syntactic rules.

4 An Architecture for Machine Translation

We now develop an architecture for machine translation. Our architecture will be general and largely language independent. We propose a hybrid model that attempts to bring the best of linguistic and statistical techniques available. We will show how this approach can lead to very efficient parsing and translation. In the next section, we will show how the development of such MT systems itself can be done quickly and efficiently.

Experience has shown that purely linguistic models and purely statistical models both have their inherent limitations. See [5] and [6] for examples of purely linguistic and purely statistical approaches to machine translation. Here we propose a hybrid model that attempts to exploit the best of both worlds and can lead to efficient creation of efficient translation systems.

We use Finite State Machines to identify word groups in a given sentence. UCSG [4] has shown that all potential word groups in a sentence can be identified in a single linear scan of the sentence and in linear time. As we have mentioned already, we do not attempt to analyse the complete internal structure of word groups. Instead we focus on merely recognizing all the word groups in a given sentence. Thus the grammars employed for this analysis are extremely simple and can be developed quickly and easily.

However, the linguistic rules deal only with ‘possibilities’ rather than probabilities and there is a tendency to generate a large number of potential word groups, of which only

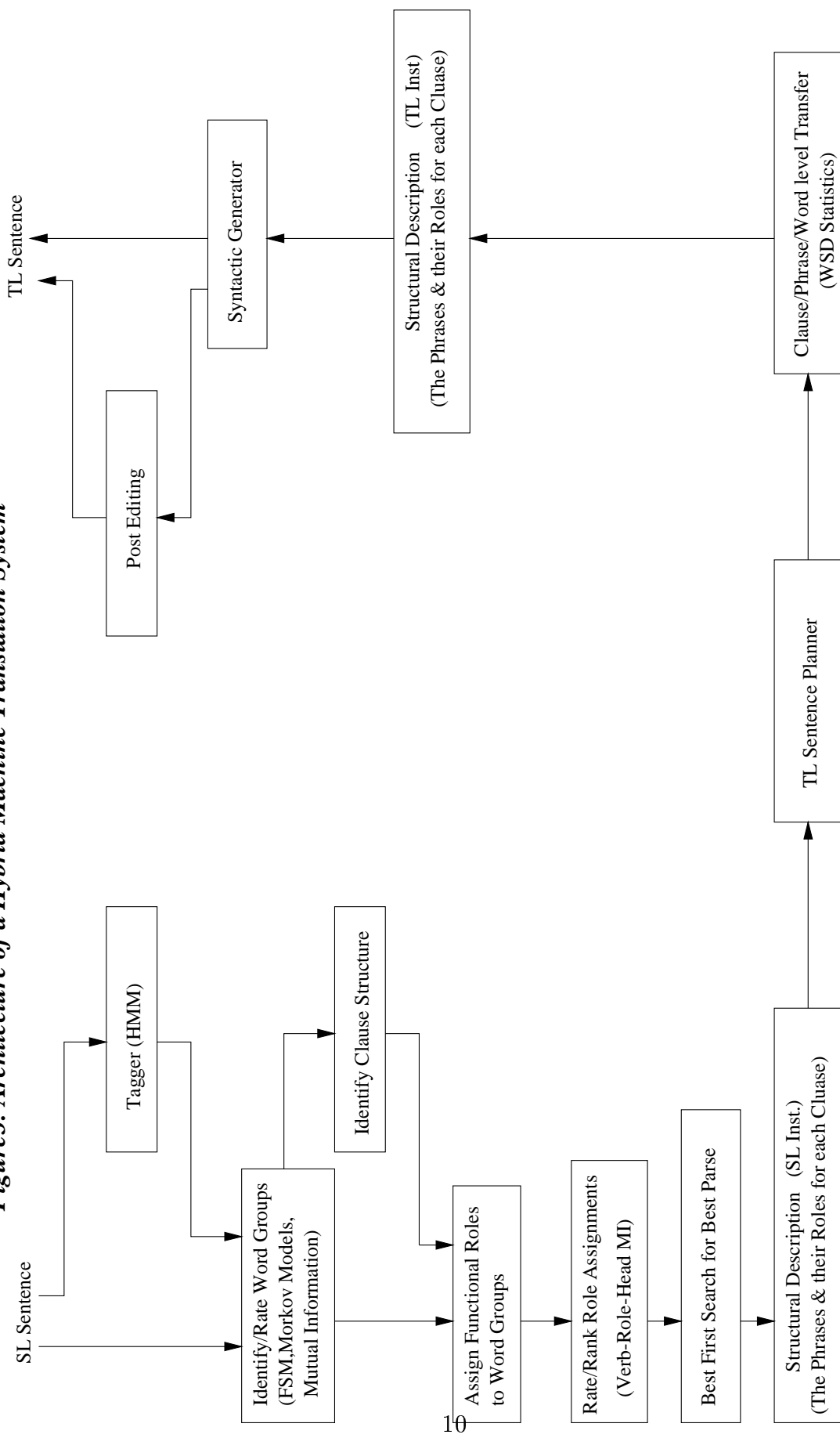
some are really correctly grouped. We resort to statistical techniques here. We can use fairly simple models such as Markov Models to rate the identified word groups and to rank them accordingly. Mutual Information may also be used for the purpose. Further, Hidden Markov Models (HMM) can be used to tag the text before word group analysis is taken up. A modified version of Viterbi algorithm is used to produce the top few best tag sequences for the given sentence.

Another place where statistical techniques can be used is in functional structure analysis. Instead of attempting to get the correct role assignment straightaway, we produce the set of all possible role assignments using simple linguistic models based on sub-categorization frames and selectional restrictions and then employ a ‘best first’ search strategy to obtain the best parse for a given sentence. This strategy allows us to retain the ability to produce all possible parses while at the same time helping us to obtain the best parses first. The quality metric employed will be based on statistics. From a parsed corpus, we obtain statistics regarding the functional roles that a phrase with a given head word can take in a clause with a given verb. These statistics summarize the selectional restrictions and preferences over the training corpus and therefore will be useful in rating the various potential role assignments.

Once a syntactic parse for the SL sentence is obtained, a sentence planner plans the overall structure of the TL sentence. Then we work one clause at a time, working from whole to part rather than from left to right, and map the various functional roles in each. This approach permits direct transfers at the sentence, clause and phrase levels to take care of frozen and idiomatic expressions, domain specific terminology, etc.

Thus the translation module itself is a fairly simple and straight forward module. The only major concern is unresolved word sense ambiguities and lexical choice. Statistical techniques are used here again to improve the performance. The overall architecture of our MT system can now be summarized in the figure below:

Figure 3: Architecture of a Hybrid Machine Translation System



5 Efficient Construction of Machine Translation systems

In this section we will address the issue of fast development of MT systems. In particular, we will see how the architecture we have proposed can lead to efficient construction of MT systems.

The architecture we have proposed is highly modular - the various modules are largely independent of one another and can thus be developed simultaneously. Separation of clause level and phrase level analysis, can make the modules much simpler. The grammar of noun groups not including relative clauses, is certainly much simpler than a grammar that mixes phrases and clauses. In fact, we have shown that Finite State Machine power is completely sufficient for phrase identification whereas at least Context Free Grammar power is required for handling nested clauses.

Usefulness of statistical techniques is limited by the availability of large scale annotated corpora. The annotation must also be dependable, lest the statistics collected therefrom be unreliable. Generating large scale annotated corpora is itself a very difficult and time consuming process. Yet, research has shown that only a small portion of the whole training corpus finally leads to the statistics collected and the rest of the corpus remains unused. Of course, it is not easy to say a-priori which parts of a corpus will be necessary and which parts are not. Naturally, the focus of recent research has been on reducing the corpus requirements. A key feature of our architecture is that the statistics required are also modularised and most of the required information can be extracted from a corpus of only simple sentences or individual clauses in a multi-clause sentence. Bootstrapping can be used to generate small but more useful training corpora. The output of the MT system under development, can be manually post-edited using a variety of powerful post-editing tools to quickly create small but very useful training corpora. This would be a parallel corpus, aligned at the clause, phrase and word levels, tagged, and parsed and thus extremely useful. The architecture proposed here is being used at the University of Hyderabad, India, for developing machine translation systems between Indian languages and English.

6 Conclusions

In this paper we have outlined the architecture of a hybrid machine translation system. The architecture is general and to a large extent language independent. We believe that this architecture can lead to very efficient machine translation systems. Efficiency is becoming increasingly more important, especially when the MT system is itself embedded into another application such as a multi-lingual information retrieval system. In particular, the following aspects of our architecture contribute to its efficiency:

- Word groups are identified right in the beginning. This reduces the effective length

of the sentence in all subsequent levels of processing. No more than FSM power is needed to recognize all potential word groups in a sentence, thereby giving linear time complexity. Tagging can further reduce lexical ambiguities.

- Potential word groups so recognized are rated and ranked. Very low probability groups can be pruned by thresholding if required. Even otherwise, ordering the word groups based on their probabilities can make the subsequent levels of processing more efficient, since a best first strategy is employed.
- UCSG theory can be incorporated to analyse clause structure of sentences. This makes the functional structure analysis local to a clause, effectively reducing the complexity of role assignment very significantly in multi-clause sentences. Clause structure analysis itself requires only CFG power and only a small number of rules.
- Role assignments are made according to straight forward linguistic rules based on sub-categorization frames and selectional restrictions. Potential role assignments are then rated and ranked using statistics based on the probability of a phrasal head taking on a specified role for a specified verb.
- Best first search strategy then permits the generation of the best parses first, while retaining the ability to generate all linguistically possible parses. Generating all possible parses and then rating or pruning would be a lot more time consuming.
- Translation can also proceed on clause, phrase and word levels. This ‘working from whole to part’ strategy itself pays big dividends in terms of simplicity and efficiency.
- Word Sense Disambiguation module can reduce lexical choices and thus lead to faster yet better translations.

The combination of linguistic and statistical techniques helps us to exploit available knowledge and resources while at the same time overcoming the limitations of speed and performance. The architecture is modular and the required linguistic and statistical components can be built independently and parallelly. Modularity helps in reducing the corpus requirements too. For building bilingual lexicons, word group rating, role assignment rating as well as Word Sense Disambiguation, statistics can be obtained from annotated corpora of simple sentences only. All inter-clause dependencies are handled separately by the clause structure grammar. Finally, bootstrapping can be employed to quickly develop practical MT systems. The system itself can be used for obtaining training data for the statistical components - the system can generate aligned, tagged, parsed and translated corpora for further enhancements. The architecture is currently being used at the University of Hyderabad, India, for building translation systems between English and Indian languages.

7 references

1. Raja1963 “Indian Theories of Meaning”, K. Kunjunni Raja, The Adyar Library and Research Centre, Adyar, Chennai, India, 1963.

2. Michael George Dyer, "In Depth Understanding", MIT Press, 1983.
3. Amorey Gethin, "Antilinguistics - A critical assessment of modern linguistic theory and practice", Intellect, 1990
4. K. Narayana Murthy, "Universal Clause Structure Grammar", PhD thesis, University of Hyderabad, 1996.
5. K. Narayana Murthy, "MAT: A Machine Assisted Translation System", proceedings of the 5th Natural Language Processing Pacific Rim Symposium, November 5-7, 1999, Beijing, China.
6. Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frderick Jelinek, John D Lafferty, Robert L Mercer and Paul S Roossin, "A Statistical Approach to Machine Translation", Computational Linguistics, Vol 16, No 2, pp 79-85, 1990