DATA MINING AND MARKER WORDS : A PSYCHOLINGUISTIC APPROACH TO MACHINE TRANSLATION

Patrick Juola
University of Colorado at Boulder

1 Data Mining

1.1 General Background

Since the very dawn of the computer age, it has been the dream of some computer scientists to be able to develop programs capable of dealing with natural language with truly human-like skill. The inherent complexity of natural language has frustrated and continues to frustrate such grandiose plans. The primary difficulties seem to lie in the difficulty of collecting and representing linguistic knowledge of sufficient depth and coverage.

As computers have developed, particularly with regards to their available disk space, memory, and speed, it has become more practical to try to develop linguistic knowledge directly from large, machine readable corpora. An early example of this sort of approach can be seen in the work of Kucera and Francis (1967) on relative word frequencies in English. The amount of data they collected for their ground-breaking work is now available in a few days to anyone with a few hundred dollars to spend on a CD-ROM or with a USENET feed.

Because of the greater availability of such corpora, research projects of this nature and scope are much more common now. For example, Martin (1994) has examined the nature of metaphor in standard written English. After examining several million words of text from the *Wall Street Journal*, he has developed a database capable of answering questions such as "what are the uses of the word 'hemorrhage' in the WSJ?" and documented the utility and productivity of such metaphors as MONEY IS BLOOD. Hearst (1994) has applied statistical techniques to the problem of identification of discourse boundaries and has demonstrated that she can reliably reproduce paragraphing boundaries by measuring intersentential word repetitions. Yarowsky (1994) has shown that word senses (such as 'bank'/financial institution vs. 'bank'/edge of a river) can be reliably disambiguated by statistical examination of the contexts in which they appear, a task similar to that performed by lexicographers and authors of dictionaries. Finally, the problem of grammar acquisition has been approached by many people, among them Lari and Young (1990), using various statistical and probabilistic formalisms.

1.2 Corpus-based Machine Translation

Machine translation is an obvious candidate for such corpus-based approaches. Consider the problem of learning the translation of a single word such as the French word *lait*/'milk'. One can envision a computer program that examines huge sets of sentences in French along with their corresponding English translations and observing that, when the word *lait* appears in the French, the word 'milk' almost always appears in the English. Similarly, if the word *lait* does not appear in the French sentence, the word 'milk' is very unlikely to appear in its English translation. Cooccurrence observations like this can be used to build a bilingual lexicon to use for translating novel sentences. The main weakness of this sort of approach is the application of the statistics to new sentences, particularly the identification and development of grammars for the source and target language.

A similar corpus-based approach was developed by Brown *et al.* (1990). Using a huge bilingual corpus, they attempted to solve the translation problem as a mapping between Markov chains, asserting that every sentence is a possible translation of every other sentence, then calculating the most probable translation from the statistics of the corpus. The limitations of such an translation approach are many and varied. There is no idea of grammar, only a simple Markov chain. There is no context-sensitivity, and no notion of selecting the most appropriate translation from a set of near-synonyms. This method, however, appears to require relatively little human expertise or time to produce its translations.

A similar approach has been used by Koncar and Guthrie (1994), who used a large neural network to infer translations between a large set of English sentences and their Serbo-Croat translations. As above, this work is weak in its ability to handle grammatical constructions, and it is very difficult to analyze the inferred functions in any sort of linguistically plausible or understandable fashion.

A final example of this sort of data mining can be seen in the work of Jones et al. (Somers *et al.* 1994; Jones and Alexa 1994) on the automatic extraction of translation functions by alignment. In its simplest terms, it uses a large German/English database as described in the toy system above. Using sophisticated techniques, it then knits together the translated fragments from many different words in a huge constraint network to produce accurate translations. In many regards, this is the most linguistically sophisticated and plausible data-mining system produced to date. It preserves, for example, the notion of compositionality in the sense of the final translation being produced by a compounding of (potentially many) translated fragments. However, the fragments themselves are not produced by any sort of a parse scheme, and are instead produced by a simple dynamic programming routine. The fragments identified are not necessarily linguistically useful or well-formed.

What is missing from these approaches is the notion of a linguistically cogent and plausible grammar of the source and target languages. By applying psycholinguistic principles to the inference task, the translation process can be simplified and made much easier to understand and modify. The following sections describe an early METLA (Machine Engineered Translation by Language Acquisition) system as developed to solve this problem.

2 The Marker Hypothesis

2.1 Statement

The main psycholinguistic underpinnings of the METLA translation system is the Marker Hypothesis as developed by Green (1979) and others (Morgan *et al.* 1989; Mori and Moeser 1983). In its simplest form, this universal states that natural languages are "marked" for grammar at surface level—that there exists in every language a small set of words or morphemes that appear in a very limited set of grammatical contexts and that can be said, in a sense, to signal that context. As an example of this principle, consider a basic sentence in English :

> The Boulder Faculty Assembly announced a list of ten faculty awards at its Thursday meeting, with more awards for excellence in teaching than expected.

In this sentence, taken at random from a Boulder newspaper, two noun phrases began with determiners, two with quantifiers, and one with a possessive pronoun. The set of determiners and possessive pronouns in English is very small (less than fifteen words, depending upon how one counts[1]), and the set of quantifiers is equally recognizable[2] Similarly, every word in this sentence ending with '-ed' is a past tense verb. The Marker Hypothesis presumes the converse of these observations, e.g. that words which end in '-ed' are very often past tense verbs, and the word 'the' usually heralds the appearance of a noun phrase. Or, more generally, that concepts and structures like these will have similar morphological or structural marking in all languages.

2.2 Psycholinguistic evidence

Proponents of the Marker Hypothesis go further, however, claiming not only that these "marker words" could signal the occurrence of particular contexts, but that they do—that marker words form an important cue to psycholinguistic processing of structure. Experiments with miniature languages have backed up this claim. When human subjects are presented with the task of learning a small artificial language from sentences in the language, they learn more accurately and faster if the artificial language has cues of the sort described above. Green (1979) showed this effect in artificial languages with and without specific marker words as attested in Japanese. Morgan *et al.* (1989) demonstrated it in languages with and without phrase-level substitutions, as of pronouns for full noun phrases. Mori and Moeser (1983) examined the effect of case marking on the pseudowords of the languages. In these and other experiments, evidence confirming the Marker Hypothesis was always found.

Other evidence for the psychological utility of marker words can be found in typological evidence. The original statement of the Marker Hypothesis was based upon the typological observation that every natural language has such constructs, whether in derivational morphology or separate marker words. Even pidgins and creoles have such constructs. For

[1] e.g., is 'thy' worth putting into a translation system?

[2] Although in theory there are an infinite number of quantifiers, words like '635' or 'heptillion' are rare and easy to process. See (Croft 1990:p. 98 et seq.) for a discussion of number markedness.

example, Slobin (1979) lists examples from a pidgin called Russenorsk. In this language, sentences tend to be very simple strings of words, without grammatical inflection. Even in this language, however, verbs are marked with a special '-om' marker, which presumably helps hearers of this language identify the basic concept expressed in a given utterance (and from that determine the appropriate roles of the other words in the sentence).

Other psycholinguistic evidence for such the Marker Hypothesis can be taken from child language acquisition. Constructs which are easily and readily marked (e.g., regular verbs) tend to be learned early and strongly, and may even override other irregular forms which have been learned by rote memorization. Slobin (1985) lists dozens of psycholinguistic principles that may describe how children focus on important bits of the language to learn. Many of these (for example, "pay attention to the ends of words") are direct descriptions of phenomena the Marker Hypothesis would predict.

Finally, there is psychological evidence about not only the universality of marker words and morphemes, but also about their cross-linguistic similarity. Certainly, such concepts as case marking, gender, and tense seem to be concepts found in a large variety of languages. Talmy (1988) suggests that, in fact, there are certain cognitive aspects or concepts that are inherently likely to be expressed grammatically (using marker morphemes or structural cues) and others that are universally expressed lexically. For example, many languages have inflections on nouns to express the number. On the other hand, there is no known language where morphemes exist to differentiate nouns referring to red objects from nouns referring to blue ones. Color, then, is not a concept expressed grammatically. The implication is not only that marker constructs exist, but that the semantic concepts and distinctions that they express tend to be expressed in other languages by other marker constructions.

2.3 Computational implications of the Marker Hypothesis

What useful properties would marker words[3] have? As described above, they may signal grammatical structure *if* the information can be properly teased out. Smith and Witten (1993) used a related hypothesis about "function words" to do inference of the grammar describing a large corpus. In their words, "the result is a relatively compact grammar that is guaranteed to cover every sentence in the source text that was used to form it" (pg. 1). In addition, they found that the inferred grammar was plausible under current syntactic theories, unlike many large-corpora projects.

Another advantage of the Marker Hypothesis, particularly with regard to translation, is the way it isolates content words, which tend to have few translations. Although the many words to many words problem in translation is difficult, most of the difficulty originates not in the translation of words like "computer" or "kidnapping" but in words like "of" or "the." Context-dependencies are typically defined in terms of the syntactic nature of the surroundings, i.e. in terms of the marker words, and can therefore be solved with a more complex theory of marker word translation rather than a more complex theory of translation in general.

[3]Or morphemes. The current work only focuses on marker *words*, but future developments (Hall *et al.* 1994) will include morphological analysis from large corpora as a part of marker identification.

2.4 Marker-normal form

How, then can the Marker Hypothesis be formally incorporated into a computational theory of language in a way that allows it to be easily used? As described above, the crucial property for this work is the existence of identifiable classes of marker words. Specifically, the formalism and system as described below assumes first that the languages of interest can be approximated by a context-free grammar (CFG), and second, that these languages can be naturally described by CFGs in *marker-normal form*, as defined below.

The computational background to this project can be summed up in the following mathematical result, reproduced here without proof from (Juola 1994:pg. 8-9)

Theorem 1 *To every CFG Γ there corresponds an equivalent grammar in marker-normal form, where every production is of one of the following forms :*

$$A \rightarrow \epsilon$$
$$A \rightarrow a$$
$$A \rightarrow A_0 a_1 A_1 a_2 A_2 \cdots$$
$$A \rightarrow a_1 A_1 a_2 A_2 \cdots$$

Furthermore, the Marker Hypothesis implies that explicitly marked grammars such as these are more psychologically plausible and thus that these grammars are likely to be more natural and understandable for human languages. In particular, natural language should tend to have relatively simple descriptions in which the set of terminal symbols that appear alone in productions is distinct from the set of terminal symbols that appear in a marking context; in other words, that the set of marker words is distinct and identifiable. The existence of marker-normal form provides a framework for attempting to solve natural language problems by focusing on the marker words. In addition, the symbolic, plausible, and understandable nature of these grammars makes it easier to incorporate other principles (such as $\bar{X}$ Theory) into the grammar.

3 Design considerations

The METLA system infers a grammar and symbolic transfer functions from an aligned bilingual corpus of sentences. More accurately, the system infers a set of parameters which collectively describe a grammar and transfer functions. These parameters, in turn, are derived from and express psycholinguistic theories and constraints. These parameters include a context-free grammar or equivalently strong formalism describing the source language, a context-dependent bilingual dictionary describing the relationships among lexical types in the two languages, and a set of permutation relations describing the necessary syntactic reconstruction to convert sentences in the source language into their translations in the target language.

The system begins with a random set of parameters describing a skeletal grammar and transfer functions. Over many (potentially millions or billions of) passes through the training corpus, the parameters are tuned to reduce the differences between the translated

source sentences (as translated by the current transfer functions) and the correct translation as given in the training corpus. The final set of tuned parameters can then be tested for generalization and/or used in a standalone translation system.

Once the system has been tuned to an appropriate set of parameters (or during the tuning phase as part of performance measurement), the parameters are used in a generalized translation function as follows. The parse formalism is applied to an appropriately sized unit of text, typically a sentence, to produce a parse tree. Each leaf of the tree is translated by looking up the appropriate translation in the bilingual dictionary, and then leaves are successively permuted and concatenated until the entire tree has been concatenated into the desired target sentence.

3.1 Source grammar

The first step in the translation process, obviously, is to come up with a description of the source sentence(s) in some form amenable to further processing. By assumption and design, this should be something psycholinguistically plausible while still being easily inferrible. In practical terms, this means a context-free grammar or an equivalently strong formalism, at a minimum.

The parsing algorithm used by METLA-1 is a direct expression of the marker-normal form mathematics developed in section 2.4. Specifically, every non-terminal symbol is associated with a production rule in a modified marker-normal form. For example, a sample rule for English might be

Sentence → NP aux V det NP

where 'det' is any of the set of {a, an, the} and 'aux' is any of the set of auxilliary verbs {be, have, will, can,...} in any of their inflected forms.

Formally, the grammar can be characterized as a fixed set of rules, numbered from zero to $N-1$. Each of these rules has a fixed fanout k of non-terminal symbols, so every rule in the grammar is of the form

$$A_i \rightarrow A_x m_{i,1} A_y \cdots m_{i,k-1} A_z$$

In this notation, each $A_?$ is a non-terminal in the set $A_0 \ldots A_{N-1}$ and each $m_{?,?}$ is a set of marker words that marks the separation between the various constituents of A_i.

Parsing is done in a rather simplistic fashion. A_0 is by fiat designated as the starting symbol of the grammar, and the training sentences are parsed in a strict top-down fashion. Each sentence is partitioned into its constituents at the appearance of the leftmost element of each marker set, in order of appearance in the rule of grammar. For the sample rule above, this would divide a sentence at the first auxilliary, and then at the first determiner following. These constituents are then recursively parsed in accordance with the single rule corresponding to their nonterminal, and so on, until the sentence has been broken down into only lexicalized items.

The final parameters thus constitute a formal description of the syntactic properties of the lexical items that can be used to parse novel sentences in preparation for the restructuring and translation phases of the process.

3.2 Syntactic reconstruction

Languages differ fundamentally in the syntactic structures that they use to represent similar semantic concepts. A single language, though, usually displays a relative regularity in its structures. The differences between these languages can be expressed as a simple permutation. For example, English is an SVO language, while Japanese is an SOV language. Assuming that the grammar developed in section 3.1 can successfully parse and and identify the two NPs and the VP from an English sentence, each of these components can be translated as a unit and their translations conjoined to form the Japanese translation. Numbering the components from left to right, the Japanese is produced by appending the first, third, and second components (after translation).

A similar permutation could be carried out at every point of application of every grammatical rule in the source grammar. By repeating this translate-permute-concatenate operation recursively, any sentence in the source language can be restructured into a corresponding target structure.

3.3 Context-dependent bilingual dictionary

To a first approximation, every bit of semantic information expressed in the source sentence must be present in the target sentence. The difficulty arises from the possibility of a different and ambiguous encoding in either or both languages. For example, the word 'that' in English can either be a demonstrative determiner or a marker for a relative clause. This lexical ambiguity does not have a similar ambiguity in French : the second would be translated as *que*. Such ambiguity makes it difficult to develop a bilingual dictionary to translate English words into their corresponding French words.

For many languages, a simple one-to-one dictionary will cover large fractions of the vocabulary. For those words with multiple translations, much of the ambiguity can be resolved by looking the context, generally speaking, in which they appear. It is relatively easy to generalize the notion of a single correspondence to multiple correspondences by developing multiple one-to-one correspondence sets and selecting among them on the basis of context.

The METLA system uses multiple dictionaries to produce a context-dependent dictionary for lexical selection. Every grammatical context carries with it information about which dictionary is to be used. Within an NP, then, the translation system will use a dictionary in which the lexical entry for 'that' is *ce*, while using a different dictionary with a different entry (*que*) when translating relative clauses.

Further sophistication has been added by the incorporation of ϵ (epsilon, or the null string) as an additional lexical type in all languages. This allows words to be deleted (translated to ϵ) in some contexts, or for ϵ to be translated to another word in specific contexts to insert words as appropriate.

3.4 Tuning by parameter optimization

Implicit in the above formalisms is the notion of describing them by parameter sets. For example, each of the several dictionaries can be seen as a function mapping words (or ϵ) to

other words or as a function mapping numerical tags to other numerical tags. Each domain element can be individually mapped and changed to fit the bilingual data. Similarly, the choice of dictionaries can be described in numerical terms—in such and such a grammatical context, use dictionary number three. The end result of such description is a large number of relatively independent and tunable parameters which collectively describe a transfer function between the source and target language.

Setting the parameters at random, of course, will typically result in complete gibberish. However, by translating the source sentences in the database and comparing the translated results with the "correct" translation also listed in the database, one can produce a measure of the relative fitness of a given parameter set. Standard optimization techniques can then be applied to maximize the fitness of the parameter set. METLA uses a standard multivariate optimization algorithm called simulated annealing (Metropolis *et al.* 1953; Kirkpatrick *et al.* 1983). See (Juola 1994) for a detailed description of the engineering of the inference.

4 Description of experiments

The standard procedure for most modern learning systems (e.g.Koncar and Guthrie 1994) is to produce two separate sets of data, a training set and a testing set. The system is trained to some criterion, usually measured either in terms of performance or else a set number of training epochs, and then the actual performance measurements are taken on novel data to which the system has not been exposed. This prevents the system from merely memorizing the input data and provides a better measure of learning performance, but also requires that the researchers acquire two sets of data. In the case of METLA, this would of course be two similar aligned corpora on the same language pair, or more simply two halves of the same corpus. Reported here are the results from experiments on the following two corpora :

The first corpus was an English→Urdu text taken from ur Rahman (1958). The corpus consisted of a vocabulary list and the set of example sentences (and their translations) taken from lesson 2, while the testing corpus was the set of exercises (which were translated by hand and confirmed by a native speaker of Urdu). Typologically, Urdu is an Indo-European language with a heavy influence from Arabic. Structurally, it has basic word order SOV, postpositions instead of prepositions, and no definite/indefinite article distinction.

The other corpus was an artificial English→French corpus designed to test the performance of the system on a small vocabulary but with greater syntactic complexity than the Urdu corpus. The training set consisted of forty-three sentences with words selected from a thirty word vocabulary. This corpus included various forms of lexical and syntactic complexity such as gender distinctions, embedded relative clauses, words with ambiguous translations, reflexive and non-reflexive verbs, and multiple subcategorizations of verbs. The testing data consisted of similar sentences produced by a different experimenter from the same vocabulary. All translations were confirmed by a native speaker. Typologically, French is an Indo-European language with the same basic word order and structure as English, but a more pronounced gender agreement system.

5 Evaluation of METLA-1

One of the difficulties involved with the development of a machine translation system is the evaluation of the end product. Is it better, for instance, to produce an ungrammatical translation that nonetheless seems to capture the meaning of what the original said, or to produce a grammatically flawless sentence that states something completely different from the original? How should the system respond to unusual, metaphoric, or ungrammatical inputs?

5.1 Black box evaluation

For many fully self-automated translation systems (e.g. Brown *et al.* 1990), the problem can be made worse by the relative opacity of the inferred translation system. There is no easy way to examine the internal workings of the algorithm to determine the nature and causes of a translation error or to identify how to repair the error. And for translation systems using Markov models (Brown *et al.* 1990) and similar oversimplified grammatical structures, it may not be possible to understand the cause of the error even after a lengthy and extensive analysis of the translation parameters, as the underlying model is too distant from people's intuitive understanding of how languages are put together.

Nonetheless, it is possible to do some sort of a black box analysis of the output of the system. Brown *et al.* , for instance, performed their analysis on the basis of hand-classification of sentences into five types, ranging from "Exact" (Identical to what the Hansard translator chose), through "Alternate" (Different phrasing but the same idea expressed), down to "Ungrammatical." This sort of hand-classification for final system evaluation is useful because it directly measures the appropriateness of the final product in a way that more automatic measures (such as diff) cannot. The METLA-1 prototype used a similar but less detailed classification. Because of the limited vocabulary and grammar in the experiments, few different grammatical ways to express the same idea were available. It was therefore more useful and appropriate to classify sentences (again by hand) into the categories "Correct," "Minor errors," and "Gibberish." The first category corresponds to "Exact," above. The third category describes sentences that were so syntactically ill-formed as to be unintelligible. The second category would be classified by Brown *et al.* (1990) sometimes as "Alternate" and sometimes as "Ungrammatical." These tend to be syntactically invalid but semantically understandable. Examples of these from the English→French experiments include deletion of sentence complementizers, deletion of reflexive particles, or gender errors.

When this sort of analysis is performed on the results of the English→Urdu experiments, the system learned the original training corpus (the example sentences from the lessons) perfectly and could reproduce it without errors. Testing on novel sentences (the exercises) revealed 72% completely correct, and only 7% translated as "gibberish." Upon further analysis (see section 5.2), the training corpus was shown to be unrepresentative of the test corpus, and in particular was missing coverage in context for several words. When the training corpus was updated to include coverage for the missing items, the system could still learn the training corpus perfectly and the percentage correct on novel items of the same forms increased to 100%.

Category	Training	Testing	Limited
Exact	61%	36%	41%
Minor	29%	21%	19%
Gibberish	10%	44%	41%

Table 1: Results from black-box analysis of French experiments

The English→French experiment, because of the higher syntactic complexity in conjunction with the limited scale of the prototype, performed less well overall. Typical performance for the system on the training corpus was approximately 61% correct. On the test data, performance was lower, with only 36% correct and a full 44% gibberish. However, when the test sentences that presented structures unrepresented in the grammar were excluded, the performance improved, up to 41% correct. Although cross-system and cross-corpus comparisons can be problematic, or even meaningless, the percentage correct for the METLA-1 system is in the approximate area of the results from (Brown *et al.* 1990), where an early version of the system was able to correctly translate 48% of the test data based on a much larger training (and testing) corpus. These results are summarized in tabular form in table 1.

Given the known structural limitations of the implementation and the small grammars that it used for these experiments, this represents a significant accomplishment in the development of a psycholinguistically plausible MT system. Perhaps equally significantly, to convert the system from one language to another required approximately an hour of human effort to type in the training data, and no system modifications. This indicates that language-independent induction of transfer functions may be a viable approach to machine translation.

5.2 White box evaluation

A major advantage of a psycholinguistically plausible approach is that, if properly done, the output of the system can be directly converted into a grammar and dictionaries for the appropriate languages. This makes it possible to directly analyze the plausibility and appropriateness of the various transfer rules and to improve them by human intervention.

For example, in the English→Urdu experiment, the training data consisted of copula-locatives ("the hat is on the chair", "the man is in the shop") and imperative sentences ("wait in the office," "send the knife to the house"). Upon examination, the word classification and translation methods make sense. For an example, one of the early experiments initially divided all sentences into two parts based on the first appearance of a determiner or preposition. This divided imperatives ("wait in the office") into their verb components followed by one or more arguments which were translated by another set of rules. The translation of the verb was permuted to follow the rest of the sentence, giving the necessary verb-final form. On the other hand, declarative sentences ("the book is on the table") are passed through this initial rule unchanged, to be divided later at 'is' into subject, verb, and location, and permuted appropriately. This sort of analysis can be carried out to any desired level of detail.

Even this simplified analysis, however, is enough to demonstrate the advantage of a psycholinguistically plausible and symbolic representation. The statement "to be divided later at 'is'" is, in point of fact, slightly inaccurate. Using the first version of the training data, the system accurately inferred that 'is' serves to mark the boundary between subject and verb. However, it also inferred (wrongly) that 'knife' and 'man' were also part of that same marker group. This resulted in a small number of incorrect translations of the testing sentences.

Further examination of the input corpus showed the reason that these errors had been made. Although the system was presented with a full vocabulary list ('man'/'admi', 'house'/'ghar', and so forth) of individual words, only a subset of those words had been presented in the context of a phrase or sentence. Although the system, then, had learned that 'man' translated to 'admi,' it had no evidence about the part of speech of 'man.' The system had no way of knowing, for example, that the word 'man' was not an alternate form of the copula. In general, the lists of marker words are obviously of one or more grammatical classes, with a few outliers that represent words that have never been seen in that context and therefore may or may not be relevant. With this observation, it became obvious that the input examples were not representative of the testing data, and that some new input was required. After adding two more sentences to provide context for these words, the percentage correct increased in later experiments to 100%.

Similar analysis can be done for the more grammatically-complex English→French experiments. Because of the greater syntactic complexity, the system as built proved to be oversimplified in several important regards and some errors were in that sense inevitable. On the other hand, the system correctly learned appropriate translation structure for a large part of the input corpus. For example, the original sentences are parsed into three pieces based upon the existence first of a verb, and then of a determiner or pronoun. Noun phrases (which begin with a determiner in the input corpus) are themselves partitioned into classes of masculine/feminine noun phrases so that the gender of the determiner is correctly set.

The major error made by the English→French system was that it found a local maximum in reusing one of the production rules. Because any translation system should allow for recursive structures ("John said that Mary told him that Susan said that ..."), the system is permitted to call rules that have already been called. The system tended to find a local maximum where the rule used to separate masculine from feminine nouns was the same rule used to parse the original sentence, and so it conflated the two categories of verbs and feminine nouns. This meant, in turn, that sentences such ase "that woman washes a car" were divided not as "(that woman) (washes) (a car)" but instead as *"(that) (woman washes) (a car)." This error could presumably be rectified by allowing the system to use more production rules, but is more appropriately solved by a better-parsing algorithm in general (as in METLA-2).

Some sample results are attached as tables 2 and 3. Each table shows a number of sample sentences (in the nearly opaque parenthesized format) along with their primary division into constituents, the translations of those constituents, and the final translation after it has been permuted and concatenated.

The errors in table 3 should be explained. First, note that the division of the third

bring the letter from the shop (bring) ((the letter) (from the shop)) (lao) ((chitthi) (dukan se)) chitthi dukan se lao
wait in the office (wait) (in the office) (thairo) (daftar men) daftar men thairo
put the box on the table (put) ((the box) (on the table)) (rakho) ((sanduq) (mez par)) sanduq mez par rakho

Table 2: Sample English→Urdu translations with partial analysis

the glass touches a car (the glass) (touches) (a car) (le verre) (touche) (une voiture) le verre touche une voiture
she washes a cat (she) (washes) (a cat) (elle) (lave) (un chat) elle lave un chat
the man that touches a car touches a glass (the man that) (touches) (a car touches a glass) (le homme qui) (touche) (une voiture touche un verre) le homme qui touche une voiture touche un verre
that man washes a car that she creates (that man) (washes) (a car that she creates) (ce homme) (lave) *(une voiture qui elle creee) *ce homme lave une voiture qui elle creee
this cat washes (this cat) (washes) () (ce chat) (lave) () *ce chat lave

Table 3: Sample English→French translations with partial analysis

sentence is incorrect—"the man that touches the car" is an entire component and the main verb of the sentence is the *second* token of 'touches.' This is an artifact of the admittedly broken METLA-1 parsing algorithm, which divides at the first appearance of a given token. That this sentence is correctly translated at all is a tribute to the remarkable structural similarity between this sentence and its French translation. The fifth sentence is an example of a so-called "reflexive" verb; the proper translation should be "ce chat se lave," where 'se' is a general pronoun meaning 'self.' In English, certain verbs can be intransitive when the subject and object of the verb are the same—for example, "I shave (myself) every morning," "I wash (up)," and so forth. Some of these verbs, in turn, *must* be expressed with the reflexive particle in French (when the English sentence is intransitive) but with an ordinary direct object when the English sentences is transitive. This leads, in turn, to an example of an oversimplification; in this case, the assumption that only one rule or permutation is necessary per given non-terminal symbol.

The fourth sentence is more interesting. The word 'qui' in the fourth example sentence is a relative pronoun used only for people (like 'who'). As an inanimate object, "a car" should have taken the relative pronoun 'que' as a translation of 'that'. However, notice should be taken of the mistake that the system did not make. The other token of 'that' in the sentence was a demonstrative determiner, which was correctly translated as 'ce', taking into account the gender of 'man'. The system correctly identified the second 'that' as a relative pronoun and not a demonstrative determiner. Similarly, the third sentence indicates an ability to distinguish between feminine nouns ("une voiture") and masculine ones ("un verre"), a relatively subtle grammatical point. These results, then, indicate an ability on the part of METLA-1 to determine remarkably small grammatical structures and to appropriately account for and to produce them as needed in the translation process.

6 Summary

Despite the evident limitations of the formalism, the results from the METLA-1 system were promising. The system demonstrated an ability to identify useful and psycholinguistically plausible structural regularities in bilingual corpora, and in particular identified such syntactic constructions as noun phrases, prepositional phrases, and verb phrases. It further identified the relationship among similar roles such as subject and object and correctly found a method of restructuring between two disparate languages.

Furthermore, the system could distinguish between multiple senses and uses of the same word(s), either within a language or across multiple languages (such as the gender distinctions in French). Finally, the system produced these results purely on the basis of examining the bilingual corpus and did not have to be specifically modified to handle a particular language category.

On the other hand, some problems were clearly apparent with the system. First and foremost, METLA-1 cannot handle multiple productions per nonterminal symbol, resulting in a tremendous loss of expressive power in the inferred grammars. Second, because of the greedy parsing scheme, the system failed to identify embedded clauses or handle many forms of recursion properly. And, finally, the system as designed cannot handle vocabularies larger than 31 words, so scalability is nearly nonexistent. The next version of the system, METLA-

2, was designed to overcome these limitations both with more general data structures and a more powerful and psycholinguistically plausible parsing formalism.

References

BROWN, PETER F., JOHN COCKE, STEPHEN A. DELLA PIETRA, VINCENT J. DELLA PIETRA, FREDRICK JELINEK, JOHN D. LAFFERTY, ROBERT L. MERCER, and PAUL S. ROOSSIN. 1990. A statistical approach to machine translation. *Computational Linguistics* 16.79–85.

CROFT, WILLIAM. 1990. *Typology and Universals*. Cambridge: Cambridge University Press.

GREEN, T. R. G. 1979. The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behavior* 18.481–96.

HALL, CHRIS, PATRICK JUOLA, and ADAM BOGGS. 1994. Morpheus : A tool for the lexical analysis of corpora for morpheme segmentation. In *Proceedings of the 1994 Mid-America Linguistics Conference*, Lawrence, Kansas.

HEARST, MARTI A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, 9–16, Las Cruces, New Mexico.

JONES, DANIEL, and MELINA ALEXA. 1994. Towards automatically aligning German compounds with English word groups in an example-based translation system. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, 66–70, Manchester, UK.

JUOLA, PATRICK. 1994. Self-organizing machine translation : Example-driven induction of transfer functions. Technical Report TR-722-94, Computer Science Department, University of Colorado. Also available as cmp-lg/9406012.

KIRKPATRICK, S., C. D. GELATT, JR., and M. VECCHI. 1983. Optimization by simulated annealing. *Science* 20.671–80.

KONCAR, NENAD, and GREGORY GUTHRIE. 1994. A natural language translation neural network. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, 71–77, Manchester, UK.

KUCERA, HENRY, and W. NELSON FRANCIS. 1967. *Computational Analysis of Present-day American English*. Providence: Brown University Press.

LARI, K., and S.J. YOUNG. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language* 4.35–56.

MARTIN, J. 1994. Metabank : A knowledge-base of metaphoric language conventions. *Computational Intelligence* 10.134–149.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A.H. TELLER, and E. TELLER. 1953. Equations of state calculations by fast computing machines. *The Journal of Chemical Physics* 21.1087–92.

MORGAN, JAMES L., RICHARD P. MEIER, and ELISSA L. NEWPORT. 1989. Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language* 28.360–74.

MORI, KAZUO, and SHANNON D. MOESER. 1983. The role of syntax markers and semantic referents in learning an artificial language. *Journal of Verbal Learning and Verbal Behavior* 22.701–18.

SLOBIN, DAN ISAAC. 1979. *Psycholinguistics*. Glenview, Ill.: Scott, Foresman, and Company, second edition.

——. 1985. Crosslinguistic evidence for the language-making capacity. In *The Cross-Linguistic Study of Language Acquisition*, ed. by Dan Isaac Slobin, volume 2 : Theoretical Issues, chapter 15, 1157–1256. 365 Broadway, Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

SMITH, TONY C., and IAN H. WITTEN. 1993. Language inference from function words. Technical Report 1993/3, University of Waikato, New Zealand.

SOMERS, HAROLD, IAN MCLEAN, and DANIEL JONES. 1994. Experiments in multilingual example-based generation. In *3rd International Conference on the Cognitive Science of Natural Language Processing (CSNLP-94)*, Dublin, Ireland.

TALMY, LEONARD. 1988. The relation of grammar to cognition. In *Topics in Cognitive Linguistics*, ed. by Brygida Rudzka-Ostyn, 165–205. Amsterdam/Philadelphia: John Benjamins Publishing Co.

UR RAHMAN, AZIZ. 1958. *Teach Yourself Urdu in Two Months*. II, K, 14/4, Nazimabad, Karachi–18, Pakistan: Azizi's Oriental Book Depot, 22nd edition.

YAROWSKY, DAVID. 1994. Decision lists for lexical ambiguity resolution : Application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, 88–95, Las Cruces, New Mexico.