Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information

Sonja Nießen*
RWTH Aachen

Hermann Ney*
RWTH Aachen

In statistical machine translation, correspondences between the words in the source and the target language are learned from parallel corpora, and often little or no linguistic knowledge is used to structure the underlying models. In particular, existing statistical systems for machine translation often treat different inflected forms of the same lemma as if they were independent of one another. The bilingual training data can be better exploited by explicitly taking into account the interdependencies of related inflected forms. We propose the construction of hierarchical lexicon models on the basis of equivalence classes of words. In addition, we introduce sentence-level restructuring transformations which aim at the assimilation of word order in related sentences. We have systematically investigated the amount of bilingual training data required to maintain an acceptable quality of machine translation. The combination of the suggested methods for improving translation quality in frameworks with scarce resources has been successfully tested: We were able to reduce the amount of bilingual training data to less than 10% of the original corpus, while losing only 1.6% in translation quality. The improvement of the translation results is demonstrated on two German-English corpora taken from the Verbmobil task and the Nespole! task.

1. Introduction

The statistical approach to machine translation has proved successful in various comparative evaluations since its revival by the work of the IBM research group more than a decade ago. The IBM group dispensed with linguistic analysis, at least in its earliest publications. Although the IBM group finally made use of morphological and syntactic information to enhance translation quality (Brown et al. 1992; Berger et al. 1996), most of today's statistical machine translation systems still consider only surface forms and use no linguistic knowledge about the structure of the languages involved.

In many applications only small amounts of bilingual training data are available for the desired domain and language pair, and it is highly desirable to avoid at least parts of the costly data collection process. The main objective of the work reported in this article is to introduce morphological knowledge in order to reduce the amount of bilingual data necessary to sufficiently cover the vocabulary expected in testing. This is achieved by explicitly taking into account the interdependencies of related inflected forms. In this work, a hierarchy of equivalence classes at different levels of abstraction is proposed. Features from those hierarchy levels are combined to form hierarchical lexicon models, which can replace the standard probabilistic lexicon used

* Lehrstuhl für Informatik VI, Computer Science Department, RWTH Aachen—University of Technology, D-52056 Aachen, Germany. E-mail: sonja.niessen@gmx.de; ney@cs.rwth-aachen.de.

in most statistical machine translation systems. Apart from the improved coverage, the proposed lexicon models enable the disambiguation of ambiguous word forms by means of annotation with morpho-syntactic tags.

1.1 Overview

The article is organized as follows. After briefly reviewing the basic concepts of the statistical approach to machine translation, we discuss the state of the art and related work as regards the incorporation of morphological and syntactic information into systems for natural language processing. Section 2 describes the information provided by morpho-syntactic analysis and introduces a suitable representation of the analyzed corpus. Section 3 suggests solutions for two specific aspects of structural difference, namely, question inversion and separated verb prefixes. Section 4 is dedicated to hierarchical lexicon models. These models are able to infer translations of word forms from the translations of other word forms of the same lemma. Furthermore, they use morpho-syntactic information to resolve categorial ambiguity. In Section 5, we describe how disambiguation between different readings and their corresponding translations can be performed when no context is available, as is typically the case for conventional electronic dictionaries. Section 6 provides an overview of our procedure for training model parameters for statistical machine translation with scarce resources. Experimental results are reported in Section 7. Section 8 concludes the presentation with a discussion of the achievements of this work.

1.2 Statistical Machine Translation

In statistical machine translation, every target language string $e_1^I = e_1 \cdots e_I$ is assigned a probability $\Pr(e_1^I)$ of being a valid word sequence in the target language and a probability $\Pr(e_1^I|f_1^J)$ of being a translation for the given source language string $f_1^J = f_1 \cdots f_J$. According to Bayes' decision rule, the optimal translation for f_1^J is the target string that maximizes the product of the target language model $\Pr(e_1^I)$ and the string translation model $\Pr(f_1^J|e_1^I)$. Many existing systems for statistical machine translation (García-Varea and Casacuberta 2001; Germann et al. 2001; Nießen et al. 1998; Och, Tillmann, and Ney 1999) implement models presented by Brown, Della Pietra, Della Pietra, and Mercer (1993): The correspondence between the words in the source and the target strings is described by alignments that assign target word positions to each source word position. The probability that a certain target language word will occur in the target string is assumed to depend basically only on the source words aligned with it.

1.3 Related Work

1.3.1 Morphology. Some publications have already dealt with the treatment of morphology in the framework of language modeling and speech recognition: Kanevsky, Roukos, and Sedivy (1997) propose a statistical language model for inflected languages. They decompose word forms into stems and affixes. Maltese and Mancini (1992) report that a linear interpolation of word n-grams, part of speech n-grams, and lemma n-grams yields lower perplexity than pure word-based models. Larson et al. (2000) apply a data-driven algorithm for decomposing compound words in compounding languages as well as for recombining phrases to enhance the pronunciation lexicon and the language model for large-vocabulary speech recognition systems.

As regards machine translation, the treatment of morphology is part of the analysis and generation step in virtually every symbolic machine translation system. For this purpose, the lexicon should contain base forms of words and the grammatical category,

subcategorization features, and semantic information in order to enable the size of the lexicon to be reduced and in order to account for unknown word forms, that is, word forms not present explicitly in the dictionary.

Today's statistical machine translation systems build upon the work of P. F. Brown and his colleagues at IBM. The translation models they presented in various papers between 1988 and 1993 (Brown et al. 1988; Brown et al. 1990; Brown, Della Pietra, Della Pietra, and Mercer 1993) are commonly referred to as IBM models 1–5, based on the numbering in Brown, Della Pietra, Della Pietra, and Mercer (1993). The underlying (probabilistic) lexicon contains only pairs of full forms. On the other hand, Brown et al. (1992) had already suggested word forms be annotated with morpho-syntactic information, but they did not perform any investigation on the effects.

1.3.2 Translation with Scarce Resources. Some recent publications, like Al-Onaizan et al. (2000), have dealt with the problem of translation with scarce resources. Al-Onaizan et al. report on an experiment involving Tetun-to-English translation by different groups, including one using statistical machine translation. Al-Onaizan et al. assume the absence of linguistic knowledge sources such as morphological analyzers and dictionaries. Nevertheless, they found that the human mind is very well capable of deriving dependencies such as morphology, cognates, proper names, and spelling variations and that this capability was finally at the basis of the better results produced by humans compared to corpus-based machine translation. The additional information results from complex reasoning, and it is not directly accessible from the full-word-form representation in the data.

This article takes a different point of view: Even if full bilingual training data are scarce, monolingual knowledge sources like morphological analyzers and data for training the target language model as well as conventional dictionaries (one word and its translation[s] per entry) may be available and of substantial usefulness for improving the performance of statistical translation systems. This is especially the case for more-inflecting major languages like German. The use of dictionaries to augment or replace parallel corpora has already been examined by Brown, Della Pietra, Della Pietra, and Goldsmith (1993) and Koehn and Knight (2001), for instance.

2. Morpho-syntactic Information

A prerequisite for the methods for improving the quality of statistical machine translation described in this article is the availability of various kinds of morphological and syntactic information. This section describes the output resulting from morpho-syntactic analysis and explains which parts of the analysis are used and how the output is represented for further processing.

2.1 Description of the Analysis Results

For obtaining the required morpho-syntactic information, the following analyzers for German and English were applied: `gertwol` and `engtwol` for lexical analysis and `gercg` and `engcg` for morphological and syntactic disambiguation. For a description of the underlying approach, the reader is referred to Karlsson (1990). Tables 1 and 2 give examples of the information provided by these tools.

2.2 Treatment of Ambiguity

The examples in Tables 1 and 2 demonstrate the capability of the tools to disambiguate among different readings: For instance, they infer that the word *wollen* is a verb in the indicative present first-person plural form. Without any context taken into account,

Table 1
Sample analysis of a German sentence. Input: *Wir wollen nach dem Abendessen nach Essen aufbrechen.* (In English: We want to start for Essen after dinner.)

Original	Base form	Tags
Wir	wir	personal-pronoun plural first nominative
wollen	wollen	verb indicative present plural first
nach	nach	preposition dative
dem	das	definite-article singular dative neuter
Abendessen	Abend#essen	noun neuter singular dative
nach	nach	preposition dative
Essen	Essen	noun name neuter singular dative
	Esse	noun feminine plural dative
	Essen	noun neuter plural dative
	Essen	noun neuter singular dative
aufbrechen	auf|brechen	verb separable infinitive

Table 2
Sample analysis of an English sentence. Input: *Do we have to reserve rooms?*.

Original	Base form	Tags
Do	do	verb present not-singular-third finite auxiliary
we	we	personal-pronoun nominative plural first subject
have	have	verb infinitive not-finite main
to	to	infinitive-marker
reserve	reserve	verb infinitive not-finite main
rooms	room	noun nominative plural object

wollen has other readings. It can even be interpreted as derived from an adjective with the meaning "made of wool." The inflected word forms on the German part of the Verbmobil (cf. Section 7.1.1) corpus have on average 2.85 readings (1.86 for the English corpus), 58% of which can be eliminated by the syntactic analyzers on the basis of sentence context.

Common bilingual corpora normally contain full sentences, which provide enough context information for ruling out all but one reading for an inflected word form. To reduce the remaining uncertainty, preference rules have been implemented. For instance, it is assumed that the corpus is correctly true-case-converted beforehand, and as a consequence, non-noun readings of uppercase words are dropped. Furthermore, indicative verb readings are preferred to subjunctive or imperative. In addition, some simple domain-specific heuristics are applied. The reading "plural of Esse" for the German word form *Essen*, for instance, is much less likely in the domain of appointment scheduling and travel arrangements than the readings "proper name of the town Essen" or the German equivalent of the English word *meal*. As can be seen in Table 3, the reduction in the number of readings resulting from these preference rules is fairly small in the case of the Verbmobil corpus.

The remaining ambiguity often lies in those parts of the information which are not used or which are not relevant to the translation task. For example, the analyzers cannot tell accusative from dative case in German, but the case information is not essential for the translation task (see also Table 4). Section 2.4 describes a method

Table 3
Resolution of ambiguity on the Verbmobil corpus.

Disambiguation	Number of readings per word form German	English
None	2.85	1.86
By context	1.20	1.02
By preference	1.19	1.02
By selecting relevant tags	1.06	1.01
By resorting to unambiguous part	1.00	1.00

for selecting morpho-syntactic tags considered relevant for the translation task, which results in a further reduction in the number of readings per word form to 1.06 for German and 1.01 for English. In these rare cases of ambiguity it is admissible to resort to the unambiguous parts of the readings, that is, to drop all tags causing mixed interpretations. Table 3 summarizes the gradual resolution of ambiguity.

The analysis of conventional dictionaries poses some special problems, because they do not provide enough context to enable effective disambiguation. For handling this special situation, dedicated methods have been implemented; these are presented in Section 5.1.

2.3 The Lemma-Tag Representation

A full word form is represented by the information provided by the morpho-syntactic analysis: from the interpretation `gehen verb indicative present first singular`, that is, the base form plus part of speech plus the other tags, the word form `gehe` can be restored. It has already been mentioned that the analyzers can disambiguate among different readings on the basis of context information. In this sense, the information inherent in the original word forms is augmented by the disambiguating analyzer. This can be useful for choosing the correct translation of ambiguous words. Of course, these disambiguation clues result in an enlarged vocabulary. The vocabulary of the new representation of the German part of the Verbmobil corpus, for example, in which full word forms are replaced by base form plus morphological and syntactic tags (**lemma-tag representation**), is one and a half times as large as the vocabulary of the original corpus. On the other hand, the information in the lemma-tag representation can be accessed gradually and ultimately reduced: For example, certain instances of words can be considered equivalent. This fact is used to better exploit the bilingual training data along two directions: detecting and omitting unimportant information (see Section 2.4) and constructing hierarchical translation models (see Section 4). To summarize, the lemma-tag representation of a corpus has the following main advantages: It makes context information locally available, and it allows information to be explicitly accessed at different levels of abstraction.

2.4 Equivalence Classes of Words with Similar Translation

Inflected word forms in the input language often contain information that is not relevant for translation. This is especially true for the task of translating from a more inflecting language like German into English, for instance: In parallel German/English corpora, the German part contains many more distinct word forms than the English part (see, for example, Table 5). It is useful for the process of statistical machine translation to define equivalence classes of word forms which tend to be translated by the same target language word: The resulting statistical translation lexicon becomes

Table 4
Candidates for equivalence classes.

Part of speech	Candidates
Noun	Gender (masculine, feminine, neuter) and case (nominative, dative, accusative)
Verb	Number (singular, plural) and person (first, second, third)
Adjective	Gender, case, and number
Number	Case

smoother, and the coverage is considerably improved. Such equivalence classes are constructed by omitting those items of information from morpho-syntactic analysis which are not relevant for translation.

The lemma-tag representation of the corpus helps to identify the unimportant information. The definition of relevant and unimportant information, respectively, depends on many factors like the languages involved, the translation direction, and the choice of the models. We detect candidates for equivalence classes of words automatically from the probabilistic lexicon trained for translation from German to English. For this purpose, those inflected forms of the same base form which result in the same translation are inspected. For each set of tags T, the algorithm counts how often an additional tag t_1 can be replaced with a certain other tag t_2 without effect on the translation. As an example, let $T =$ 'blau-adjective', $t_1 =$'masculine' and $t_2 =$'feminine'. The two entries ('blau-adjective-*masculine*'|'blue') and ('blau-adjective-*feminine*'|'blue') are hints for detecting gender as nonrelevant when translating adjectives into English. Table 4 lists some of the most frequently identified candidates to be ignored while translating: The gender of nouns is irrelevant for their translation (which is straightforward, as the gender of a noun is unambiguous), as are the cases nominative, dative, accusative. (For the genitive forms, the translation in English differs.) For verbs the candidates number and person were found: The translation of the first-person singular form of a verb, for example, is often the same as the translation of the third-person plural form. Ignoring (dropping) those tags most often identified as irrelevant for translation results in the building of equivalence classes of words. Doing so results in a smaller vocabulary, one about 65.5% the size of the vocabulary of the full lemma-tag representation of the Verbmobil corpus, for example—it is even smaller than the vocabulary of the original full-form corpus.

The information described in this section is used to improve the quality of statistical machine translation and to better exploit the available bilingual resources.

3. Treatment of Structural Differences

Difference in sentence structure is one of the main sources of errors in machine translation. It is thus promising to "harmonize" the word order in corresponding sentences. The presentation in this section focuses on the following aspects: question inversion and separated verb prefixes. For a more detailed discussion of restructuring for statistical machine translation the reader is referred to Nießen and Ney (2000, 2001).

3.1 Question Inversion

In many languages, the sentence structure of questions differs from the structure in declarative sentences in that the order of the subject and the corresponding finite verb is inverted. From the perspective of statistical translation, this behavior has some dis-

advantages: The algorithm for training the parameters of the target language model $\Pr(e_1^I)$, which is typically a standard n-gram model, cannot deduce the probability of a word sequence in an interrogative sentence from the corresponding declarative form. The same reasoning is valid for the lexical translation probabilities of multiword-phrase pairs. To harmonize the word order of questions with the word order in declarative sentences, the order of the subject (including the appendant articles, adjectives etc.) and the corresponding finite verb is inverted. In English questions supporting *do*s are removed. The application of the described preprocessing step in the bilingual training corpus implies the necessity of restoring the correct forms of the translations produced by the machine translation algorithm. This procedure was suggested by Brown et al. (1992) for the language pair English and French, but they did not report on experimental results revealing the effect of the restructuring on the translation quality.

3.2 Separated Verb Prefixes

German prefix verbs consist of a main part and a detachable prefix, which can be shifted to the end of the clause. For the automatic alignment process, it is often difficult to associate one English word with more than one word in the corresponding German sentence, namely, the main part of the verb and the separated prefix. To solve the problem of separated prefixes, all separable word forms of verbs are extracted from the training corpus. The resulting list contains entries of the form `prefix|main`. In all clauses containing a word matching a main part and a word matching the corresponding prefix part occurring at the end of the clause, the prefix is prepended to the beginning of the main part.

4. Hierarchical Lexicon Models

In general, the probabilistic lexicon resulting from training the translation model contains all word forms occurring in the training corpus as separate entries, not taking into account whether or not they are inflected forms of the same lemma. Bearing in mind that typically more than 40% of the word forms are seen only once in training (see, for example, Table 5), it is obvious that for many words, learning the correct translations is difficult. Furthermore, new input sentences are expected to contain unknown word forms, for which no translation can be retrieved from the lexicon. This problem is especially relevant for more-inflecting languages like German: Texts in German contain many more distinct word forms than their English translations. Table 5 also reveals that these words are often generated via inflection from a smaller set of base forms.

4.1 A Hierarchy of Equivalence Classes of Inflected Word Forms

As mentioned in Section 2.3, the lemma-tag representation of the information from morpho-syntactic analysis makes it possible to gradually access information with different grades of abstraction. Consider, for example, the German verb form `ankomme`, which is the indicative present first-person singular form of the lemma `ankommen` and can be translated into English by `arrive`. The lemma-tag representation provides an "observation tuple" consisting of

- the original full word form (e.g., `ankomme`),
- morphological and syntactic tags (part of speech, tense, person, case, ...) (e.g., `verb, indicative, present tense, 1st person singular`), and

- the base form (e.g., `ankommen`).

In the following, $t_0^i = t_0, \ldots, t_i$ denotes the representation of a word where the base form t_0 and i additional tags are taken into account. For the example above, $t_0 =$ `ankommen`, $t_1 =$ `verb`, and so on. The hierarchy of equivalence classes $\mathcal{F}_0, \ldots, \mathcal{F}_n$ is as follows:

$$\begin{aligned}
\mathcal{F}_n = \mathcal{F}(t_0^n) &= \texttt{ankommen verb indicative present singular 1} \\
\mathcal{F}_{n-1} = \mathcal{F}(t_0^{n-1}) &= \texttt{ankommen verb indicative present singular} \\
\mathcal{F}_{n-2} = \mathcal{F}(t_0^{n-2}) &= \texttt{ankommen verb indicative present} \\
&\vdots \\
\mathcal{F}_0 = \mathcal{F}(t_0) &= \texttt{ankommen}
\end{aligned}$$

where n is the maximum number of morpho-syntactic tags. The mapping from the full lemma-tag representation back to inflected word forms is generally unambiguous; thus $\mathcal{F}_n$ contains only one element, namely, `ankomme`. $\mathcal{F}_{n-1}$ contains the forms `ankomme`, `ankommst`, and `ankommt`; in $\mathcal{F}_{n-2}$ the number (`singular` or `plural`) is ignored, and so on. The largest equivalence class contains all inflected forms of the base form `ankommen`.[1] Section 4.2 introduces the concept of *combining* information at different levels of abstraction.

4.2 Log-Linear Combination

In modeling for statistical machine translation, a hidden variable a_1^J, denoting the hidden alignment between the words in the source and target languages, is usually introduced into the string translation probability:

$$\Pr(f_1^J|e_1^I) = \sum_{a_1^J} \Pr(f_1^J, a_1^J|e_1^I) = \sum_{a_1^J} \Pr(a_1^J|e_1^I) \cdot \Pr(f_1^J|a_1^J, e_1^I) \quad (1)$$

In the following, $T_j = \left(t_0^n\right)_j$ denotes the lemma-tag representation of the jth word in the input sentence. The sequence T_1^J stands for the sequence of readings for the word sequence f_1^J and can be introduced as a new hidden variable:

$$\Pr(f_1^J|a_1^J, e_1^I) = \sum_{T_1^J} \Pr(f_1^J, T_1^J|a_1^J, e_1^I) \quad (2)$$

which can be decomposed into

$$\Pr(f_1^J|a_1^J, e_1^I) = \sum_{T_1^J} \prod_{j=1}^{J} \Pr(f_j, T_j|f_1^{j-1}, T_1^{j-1}, a_1^J, e_1^I) \quad (3)$$

1 The order of omitting tags can be defined in a natural way depending on the part of speech. In principle this decision can also be left to the maximum-entropy training, when features for all possible sets of tags are defined, but this would cause the number of parameters to explode. As the experiments in this work have been carried out only with up to three levels of abstraction as defined in Section 4.2, the set of tags of the intermediate level is fixed, and thus the priority of the tags needs not be specified. The relation between this equivalence class hierarchy and the suggestions in Section 2.4 is clear: Choosing candidates for morpho-syntactic tags not relevant for translation amounts to fixing a level in the hierarchy. This is exactly what has been done to define the intermediate level in Section 4.2.

Let $\mathcal{T}(f_j)$ be the set of interpretations which are regarded valid readings of f_j by the morpho-syntactic analyzers on the basis of the whole-sentence context f_1^J. We assume that the probability functions defined above yield zero for all other readings, that is, when $T_j \notin \mathcal{T}(f_j)$. Under the usual independence assumption, which states that the probability of the translation of words depends only on the identity of the words associated with each other by the word alignment, we get

$$\Pr(f_1^J|a_1^J, e_1^I) = \sum_{\substack{T_1^J \\ T_j \in \mathcal{T}(f_j)}} \prod_{j=1}^{J} p(f_j, T_j|e_{a_j}) \tag{4}$$

As has been argued in Section 2.2, the number of readings $|\mathcal{T}(f_j)|$ per word form can be reduced to one for the tasks for which experimental results are reported here.

The elements in equation (4) are the joint probabilities $p(f, T|e)$ of f and the readings T of f given the target language word e. The maximum-entropy principle recommends choosing for p the distribution which preserves as much uncertainty as possible in terms of maximizing the entropy, while requiring p to satisfy constraints which represent facts known from the data. These constraints are encoded on the basis of feature functions $h_m(x)$, and the expectation of each feature h_m over the model p is required to be equal to the observed expectation. The maximum-entropy model can be shown to be unique and to have an exponential form involving a weighted sum over the feature functions h_m (Ratnaparkhi 1997). In equation (5), the notation t_0^n is used again for the lemma-tag representation of an input word (this was denoted by T in equations (2)–(4) for notational simplicity):

$$p(f, T|e) = p_\Lambda(f, t_0^n|e) = \frac{\exp\left[\sum_m \lambda_m h_m(e, f, t_0^n)\right]}{\sum_{\tilde{f}, \tilde{t}_0^n} \exp\left[\sum_m \lambda_m h_m(e, \tilde{f}, \tilde{t}_0^n)\right]} \tag{5}$$

where $\Lambda = \{\lambda_m\}$ is the set of model parameters with one weight λ_m for each feature function h_m. These model parameters can be trained using converging iterative training procedures like the ones described by Darroch and Ratcliff (1972) or Della Pietra, Della Pietra, and Lafferty (1995).

In the experiments presented in this article, the sum over the word forms $\tilde{f}$ and the readings $\tilde{t}_0^n$ in the denominator of equation (5) is restricted to the readings of word forms having the same base form and partial reading as a word form f'' aligned at least once to e.

The new lexicon model $p_\Lambda(f, t_0^n|e)$ can now replace the usual lexicon model $p(f|e)$, over which it has the following main advantages:

- The decomposition of the modeled events into feature functions allows meaningful probabilities to be provided for word forms that have not occurred during training as long as the feature functions involved are well-defined. (See also the argument later in the article and the definition of first-level and second-level feature functions presented in Section 4.2.1.)
- Introducing the hidden variable $T = t_0^n$ and constraining the lexicon probability to be zero for interpretations considered nonvalid readings of

f (that is, for $t_0^n \notin \mathcal{T}(f)$) amounts to making context information from the complete sentence f_1^J locally available: The sentence context was taken into account by the morpho-syntactic analyzer, which chose the valid readings $\mathcal{T}(f)$.

4.2.1 Definition of Feature Functions. There are numerous possibilities for defining feature functions. We do not need to require that they all have the same parametric form or that the components be disjoint and statistically independent. Still, it is necessary to restrict the number of parameters so that optimizing them is practical. We used the following types of feature functions, which have been defined on the basis of the lemma-tag representation (see Section 2.3):

First level: $m = \{L, \tilde{e}\}$, where L is the base form:

$$h^1_{L,\tilde{e}}(e, f, t_0^n) = \begin{cases} 1 & \text{if } e = \tilde{e} \text{ and } t_0 = L \text{ and } f \in \mathcal{F}(t_0^n) \quad (*) \\ 0 & \text{otherwise} \end{cases}$$

Second level: $m = \{T, L, \tilde{e}\}$, with subsets T of cardinality $\leq n$ of morpho-syntactic tags considered relevant (see Section 2.4 for a description of the detection of relevant tags):

$$h^2_{T,L,\tilde{e}}(e, f, t_0^n) = \begin{cases} 1 & \text{if } (*) \text{ and } T \subseteq t_1^n \quad (**) \\ 0 & \text{otherwise} \end{cases}$$

Third level: $m = \{F, T, L, \tilde{e}\}$, with the fully inflected original word form F:

$$h^3_{F,T,L,\tilde{e}}(e, f, t_0^n) = \begin{cases} 1 & \text{if } (**) \text{ and } F = f \\ 0 & \text{otherwise} \end{cases}$$

In terms of the hierarchy introduced in Section 4.1, this means that information at three different levels in the hierarchy is combined. The subsets T of relevant tags mentioned previously fix the intermediate level.[2] This choice of the types of features as well as the choice of the subsets T is reasonable but somewhat arbitrary. Alternatively one can think of defining a much more general set of features and applying some method of feature selection, as has been done, for example, by Foster (2000), who compared different methods for feature selection within the task of translation modeling for statistical machine translation. Note that the log-linear model introduced here uses one parameter per feature. For the Verbmobil task, for example, there are approximately 162,000 parameters: 47,800 for the first-order features, 55,700 for the second-order features, and 58,500 for the third-order features. No feature selection or threshold was applied: All features seen in training were used.

4.2.2 Training Procedure. The overall process of training and testing with hierarchical lexicon models is depicted in Figure 1. This figure includes the possibility of using restructuring operations as suggested in Section 3 in order to deal with structural differences between the languages involved. This can be especially advantageous in the case of multiword phrases which jointly fulfill a syntactic function: Not merging them

2 Of course, there is not only one set of relevant tags, but at least one per part of speech. In order to keep the notation as simple as possible, this fact is not accounted for in the formulas and the textual descriptions.

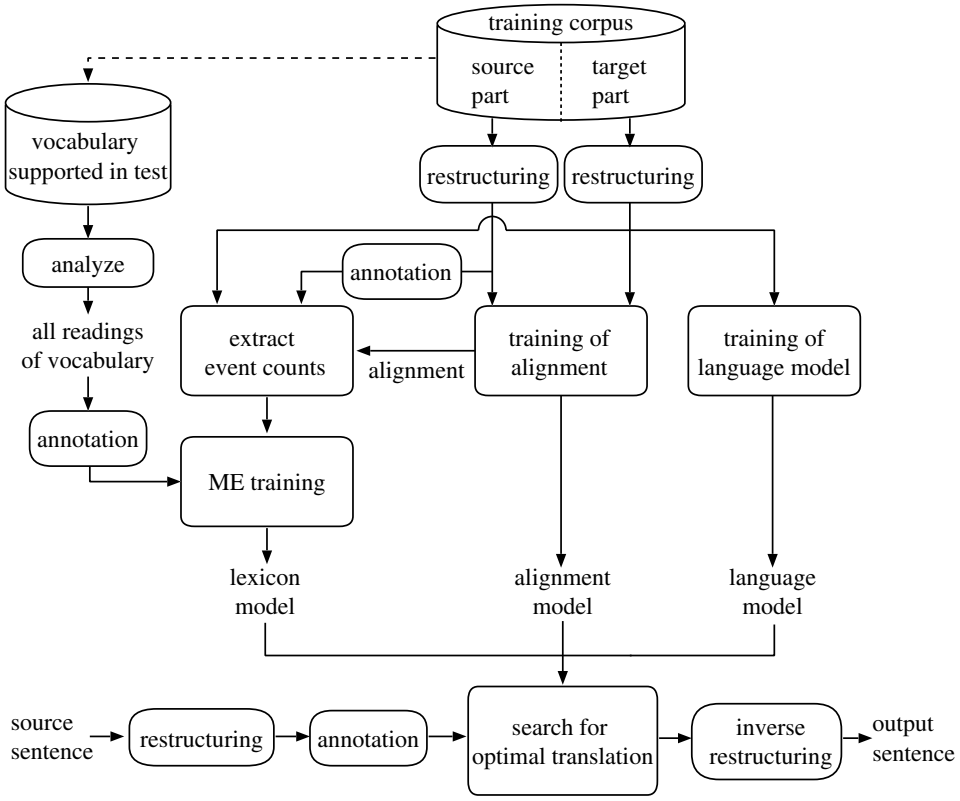

Figure 1
Training and test with hierarchical lexicon. "(Inverse) restructuring," "analyze," and "annotation" all require morpho-syntactic analysis of the transformed sentences.

would raise the question of how to distribute the syntactic tags which have been associated with the whole phrase. In Section 5.2 we describe a method of learning multi-word phrases using conventional dictionaries. The alignment on the training corpus is trained using the original source language corpus containing inflected word forms. This alignment is then used to count the co-occurrences of the annotated "words" in the lemma-tag representation of the source language corpus with the words in the target language corpus. These event counts are used for the maximum-entropy training of the model parameters Λ.

The probability mass is distributed over (all readings of) the source language word forms to be supported for test (not necessarily restricted to those occurring during training). The only precondition is that the firing features for these unseen events are known. This "vocabulary supported in test," as it is called in Figure 1, can be a predefined closed vocabulary, as is the case in Verbmobil, in which the output of a speech recognizer with limited output vocabulary is to be translated. In the easiest case it is identical to the vocabulary found in the source language part of the training corpus. The other extreme would be an extended vocabulary containing all automatically generated inflected forms of all base forms occurring in the training corpus. This vocabulary is annotated with morpho-syntactic tags, ideally under consideration of all possible readings of all word forms.

To enable the application of the hierarchical lexicon model, the source language input sentences in test have to be analyzed and annotated with their lemma-tag representation before the actual translation process. So far, the sum over the readings in equation (4) has been ignored, because when the techniques for reducing the amount of ambiguity described in Section 2.2 and the disambiguated conventional dictionaries resulting from the approach presented in Section 5.1 are applied, there remains almost always only one reading per word form.

5. Conventional Dictionaries

Conventional dictionaries are often used as additional evidence to better train the model parameters in statistical machine translation. The expression *conventional dictionary* here denotes bilingual collections of word or phrase pairs predominantly collected "by hand," usually by lexicographers, as opposed to the probabilistic lexica, which are learned automatically. Apart from the theoretical problem of how to incorporate external dictionaries in a mathematically sound way into a statistical framework for machine translation (Brown, Della Pietra, Della Pietra, and Goldsmith 1993) there are also some pragmatic difficulties: As discussed in Section 2.2, one of the disadvantages of these conventional dictionaries as compared to full bilingual corpora is that their entries typically contain single words or short phrases on each language side. Consequently, it is not possible to distinguish among the translations for different readings of a word. In normal bilingual corpora, the words can often be disambiguated by taking into account the sentence context in which they occur. For example, from the context in the sentence *Ich werde* die *Zimmer buchen*, it is possible to infer that *Zimmer* in this sentence is plural and has to be translated by *rooms* in English, whereas the correct translation of *Zimmer* in the sentence *Ich hätte gerne* ein *Zimmer* is the singular form *room*. The dictionary used by our research group for augmenting the bilingual data contains two entries for *Zimmer*: ('Zimmer'|'room') and ('Zimmer'|'rooms').

5.1 Disambiguation without Context

The approach described in this section is based on the observation that in many of the cases of ambiguous entries in dictionaries, the second part of the entry—that is, the other-language side—contains the information necessary to decide upon the interpretation. In some other cases, the same kind of ambiguity is present in both languages, and it would be possible and desirable to associate the (semantically) corresponding readings with one another. The method proposed here takes advantage of these facts in order to disambiguate dictionary entries.

Figure 2 sketches the procedure for the disambiguation of a conventional dictionary D. In addition to D, a bilingual corpus C_1 of the same language pair is required to train the probability model for tag sequence translations. The word forms in C_1 need not match those in D. C_1 is not necessarily the training corpus for the translation task in which the disambiguated version of D will be used. It does not even have to be taken from the same domain.

A word alignment between the sentences in C_1 is trained with some automatic alignment algorithm. Then the words in the bilingual corpus are replaced by a reduced form of their lemma-tag representation, in which *only a subset of their morpho-syntactic tags* is retained—even the base form is dropped. The remaining subset of tags, in the following denoted by T_f for the source language and T_e for the target language, consists of tags considered relevant for the task of aligning corresponding readings. This is not necessarily the same set of tags considered relevant for the task of *translation* which was used, for example, to fix the intermediate level for the log-linear lexicon

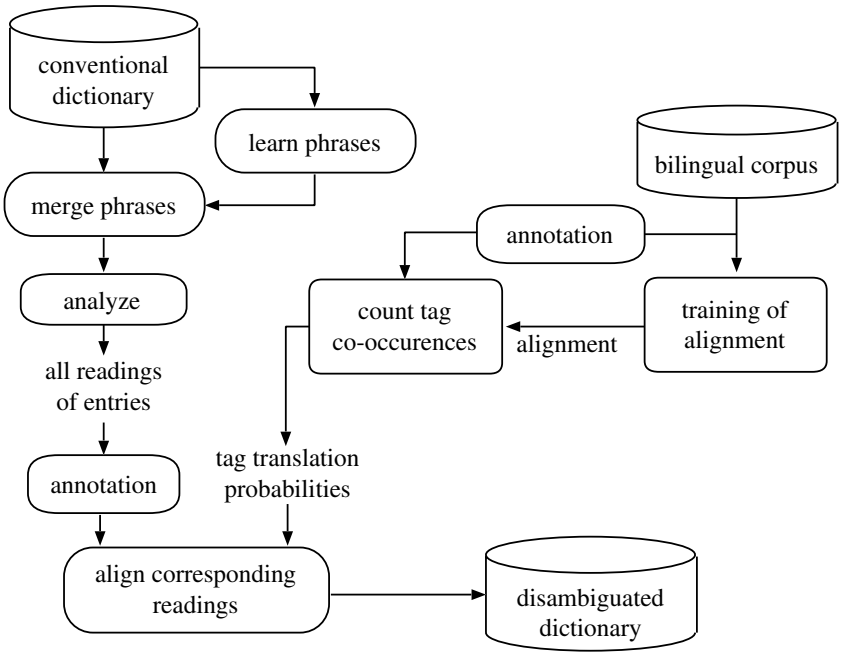

Figure 2
Disambiguation of conventional dictionaries. "Learn phrases," "analyze," and "annotation" require morpho-syntactic analysis of the transformed sentences.

combination in Section 4.2.1. In the case of the Verbmobil corpus, the maximum length of a tag sequence is five.

The alignment is used to count the frequency of a certain tag sequence $\mathbf{t}_f$ in the source language to be associated with another tag sequence $\mathbf{t}_e$ in the target language and to compute the **tag sequence translation probabilities** $p(\mathbf{t}_f|\mathbf{t}_e)$ as relative frequencies. For the time being, these tag sequence translation probabilities associate readings of *words* in one language with readings of *words* in the other language: Multiword sequences are not accounted for.

To alleviate this shortcoming it is possible and advisable to automatically detect and merge multiword phrases. As will be described in Section 5.2, the conventional bilingual dictionary itself can be used to learn and validate these phrases. The resulting multiword phrases P_e for the target language and P_f for the source language are afterwards concatenated within D to form entries consisting of pairs of "units."

The next step is to analyze the word forms in D and generate all possible readings of all entries. It is also possible to ignore those readings that are considered unlikely for the task under consideration by applying the domain-specific preference rules proposed in Section 2.2. The process of generating all readings includes replacing word forms with their lemma-tag representation, which is thereafter reduced by dropping all morpho-syntactic tags not contained in the tag sets T_f and T_e.

Using the tag sequence translation probabilities $p(\mathbf{t}_f|\mathbf{t}_e)$, the readings in one language are aligned with readings in the other language. These alignments are applied to the full lemma-tag representation (not only tags in T_f and T_e) of the expanded dictionary containing one entry per reading of the original word forms. The highest-ranking aligned readings according to $p(\mathbf{t}_f|\mathbf{t}_e)$ for each lemma are preserved.

The resulting disambiguated dictionary contains two entries for the German word *Zimmer*: (‘Zimmer-noun-sg.’|‘room-noun-sg.’) and (‘Zimmer-noun-pl.’|‘room-noun-pl.’). The target language part is then reduced to the surface forms: (‘Zimmer-noun-sg.’|‘room’) and (‘Zimmer-noun-pl.’|‘rooms’). Note that this augmented dictionary, in the following denoted by D', has more entries than D as a result of the step of generating all readings. The two entries (‘beabsichtigt’|‘intends’) and (‘beabsichtigt’|‘intended’), for example, produce three new entries: (‘beabsichtigt-verb-ind.-pres.-sg.-3rd’|‘intends’), (‘beabsichtigt-verb-past-part.’|‘intended’), and (‘beabsichtigt-adjective-pos.’|‘intended’).

5.2 Multiword Phrases

Some recent publications deal with the automatic detection of multiword phrases (Och and Weber 1998; Tillmann and Ney 2000). These methods are very useful, but they have one drawback: They rely on sufficiently large training corpora, because they detect the phrases from automatically learned word alignments. In this section a method for detecting multiword phrases is suggested which merely requires monolingual syntactic analyzers and a conventional dictionary.

Some multiword phrases which jointly fulfill a syntactic function are provided by the analyzers. The phrase *irgend etwas* (‘anything’), for example, may form either an indefinite determiner or an indefinite pronoun. irgend=etwas is merged by the analyzer in order to form one single vocabulary entry. In the German part of the Verbmobil training corpus 26 different, nonidiomatic multiword phrases are merged, while there are 318 phrases suggested for the English part. In addition, syntactic information like the identification of infinitive markers, determiners, modifying adjectives (for example, *single* room), premodifying adverbials (*more* comfortable), and premodifying nouns (*account* number) are used for detecting multiword phrases. When applied to the English part of the Verbmobil training corpus, these hints suggest 7,225 different phrases.

Altogether, 26 phrases for German and about 7,500 phrases for English are detected in this way. It is quite natural that there are more multiword phrases found for English, as German, unlike English, uses compounding. But the experiments show that it is not advantageous to use all these phrases for English. Electronic dictionaries can be useful for detecting those phrases which are important in a statistical machine translation context: A multiword phrase is considered useful if it is translated into a single word or a distinct multiword phrase (suggested in a similar way by syntactic analysis) in another language. There are 290 phrases chosen in this way for the English language.

6. Overall Procedure for Training with Scarce Resources

Taking into account the interdependencies of inflected forms of the same base form is especially relevant when inflected languages like German are involved and when training data are sparse. In this situation many of the inflected word forms to account for in test do not occur during training. Sparse bilingual training data also make additional conventional dictionaries especially important. Enriching the dictionaries by aligning corresponding readings is particularly useful when the dictionaries are used in conjunction with a hierarchical lexicon, which can access the information necessary to distinguish readings via morpho-syntactic tags. The restructuring operations described in Section 3 also help in coping with the data sparseness problem, because they make corresponding sentences more similar. This section proposes a procedure for combining all these methods in order to improve the translation quality despite sparseness of data. Figure 3 sketches the proposed procedure.

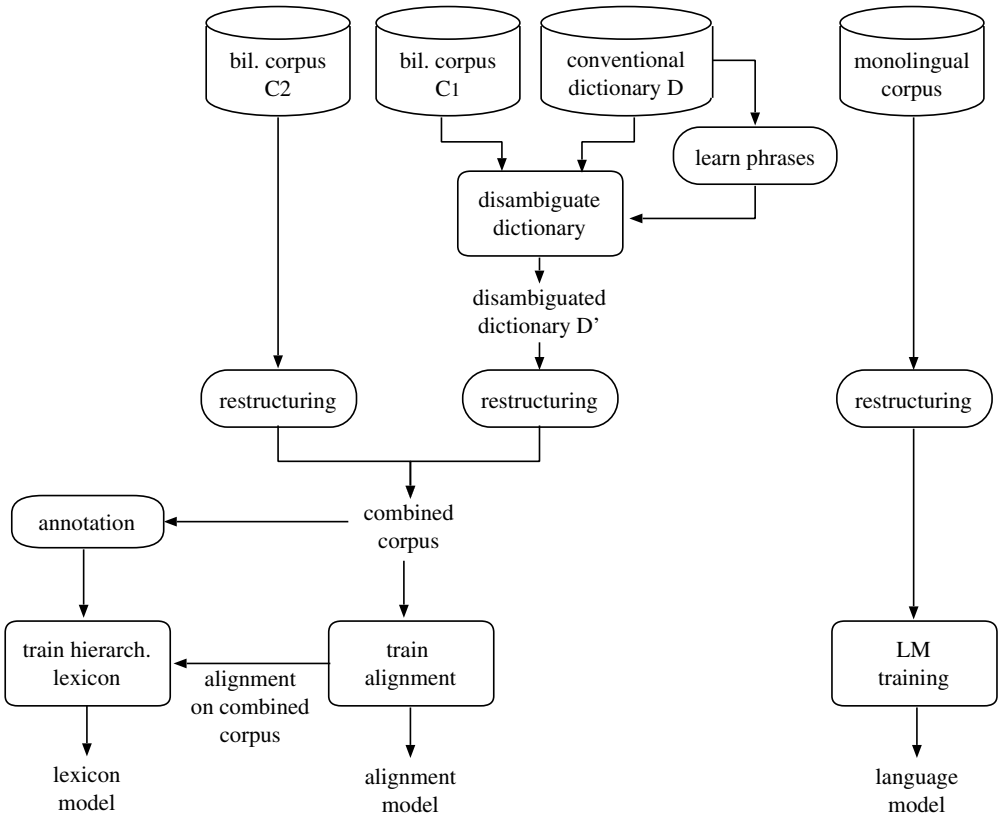

Figure 3
Training with scarce resources. "Restructuring," "learn phrases," and "annotation" all require morpho-syntactic analysis of the transformed sentences.

Two different bilingual corpora C_1 and C_2, one monolingual target language corpus, and a conventional bilingual dictionary D can contribute in various ways to the overall result. It is important to note here that C_1 and C_2 can, but need not, be distinct, and that the monolingual corpus can be identical to the target language part of C_2. Furthermore these corpora can be taken from different domains, and C_1 can be (very) small. Only C_2 has to represent the domain and the vocabulary for which the translation system is built, and only the size of C_2 and the monolingual corpus have a substantial effect on the translation quality. It is interesting to note, though, that a basic statistical machine translation system with an accuracy near 50% can be built *without any* domain-specific bilingual corpus C_2, solely on the basis of a disambiguated dictionary and the hierarchical lexicon models, as Table 9 shows.

- In the first step, multiword phrases are learned and validated on the dictionary D in the way described in Section 5.2. These multiword phrases are concatenated in D. Then an alignment is trained on the first bilingual corpus C_1. On the basis of this alignment, the tag sequence translation probabilities which are needed to align corresponding readings in the dictionary are extracted, as proposed in Section 5.1. The result of this step is an expanded and disambiguated dictionary D'. For this purpose, C_1 does not have to cover the vocabulary of D. Besides C_1

can be comparatively small, given the limited number of tag sequence pairs $(\mathbf{t}_f|\mathbf{t}_e)$ for which translation probabilities must be provided: In the Verbmobil training corpus, for example, there are only 261 different German and 110 different English tag sequences.

- In the next step, the second bilingual corpus C_2 and D' are combined, and a word alignment A for both is trained. C_2, D', and A are presented as input to the maximum-entropy training of a hierarchical lexicon model as described in Section 4.2.
- The language model can be trained on a separate monolingual corpus. As monolingual data are much easier and cheaper to compile, this corpus might be (substantially) larger than the target language part of C_2.

7. Experimental Results

7.1 The Tasks and the Corpora

Tests were carried out on Verbmobil data and on Nespole! data. As usual, the sentences from the test sets were not used for training. The training corpora were used for training the parameters of IBM model 4.

7.1.1 Verbmobil. Verbmobil was a project for automatic translation of spontaneously spoken dialogues. A detailed description of the statistical translation system within Verbmobil is given by Ney et al. (2000) and by Och (2002). Table 5 summarizes the characteristics of the English and German parallel corpus used for training the parameters of IBM model 4. A conventional dictionary complements the training corpus (see Table 6 for the statistics). The vocabulary in Verbmobil was considered closed: There are official lists of word forms which can be produced by the speech recognizers. Such lists exist for German and English (see Table 7). Table 8 lists the characteristics of the two test sets `Test` and `Develop` taken from the end-to-end evaluation in Verbmobil, the development part being meant to tune system parameters on a held-out corpus different from the training as well as the test corpus. As no parameters are optimized on the development set for the methods described in this article, most of the experiments were carried out on a joint set containing both test sets.

Table 5
Statistics of corpora for training: Verbmobil and Nespole! Singletons are types occurring only once in training.

	Verbmobil		Nespole!	
	English	German	English	German
Number of sentences	58,073	58,073	3,182	3,182
Number of *distinct* sentences	57,731	57,771	1,758	1,767
Number of running word forms	549,921	519,523	15,568	14,992
Number of running word forms without punctuation	453,612	418,974	12,461	11,672
Number of word forms	4,673	7,940	1,034	1,363
Number of singleton word forms	1,698	3,453	403	641
Number of base forms	3,639	6,063	1,072	870
Number of singleton base forms	1,236	2,546	461	326

Table 6
Conventional dictionary used to complement the training corpus.

	English	German
Number of entries	10,498	10,498
Number of running word forms	15,305	12,784
Number of word forms	5,161	7,021
Number of base forms	3,666	5,479

Table 7
The official vocabularies in Verbmobil.

	English	German
Number of word forms	6,871	10,157
Number of base forms	3,268	6,667

Table 8
Statistics for the test sets for German to English translation: Verbmobil Eval-2000 (`Test` and `Develop`) and Nespole!

	Verbmobil		Nespole!
	`Test`	`Develop`	
Number of sentences	251	276	70
Number of running word forms in German part	2,628	3,159	456
Number of word forms in German part	429	434	180
Trigram LM perplexity of reference translation	30.5	28.1	76.9

7.1.2 Nespole!. Nespole! is a research project that ran from January 2000 to June 2002. It aimed to provide multimodel support for negotiation (Nespole! 2000; Lavie et al. 2001). Table 5 summarizes the corpus statistics of the Nespole! training set. Table 8 provides the corresponding figures for the test set used in this work.

7.2 The Translation System

For testing we used the alignment template translation system, described in Och, Tillmann, and Ney (1999). Training the parameters for this system entails training of IBM model 4 parameters in both translation directions and combining the resulting alignments into one symmetrized alignment. From this symmetrized alignment, the lexicon probabilities as well as the so-called alignment templates are extracted. The latter are translation patterns which capture phrase-level translation pairs.

7.3 Performance Measures

The following evaluation criteria were used in the experiments:

BLEU (Bilingual Evaluation Understudy): This score, proposed by Papineni et al. (2001), is based on the notion of modified n-gram precision, with $n \in \{1, \ldots, 4\}$: All candidate unigram, bigram, trigram, and four-gram counts are collected and clipped against their corresponding maximum reference counts. The reference n-gram counts are calculated on a corpus

of reference translations for each input sentence. The clipped candidate counts are summed and normalized by the total number of candidate *n*-grams. The geometric mean of the modified precision scores for a test corpus is calculated and multiplied by an exponential brevity penalty factor to penalize too-short translations. BLEU is an accuracy measure, while the others are error measures.

m-WER (multireference word error rate): For each test sentence there is a set of reference translations. For each translation hypothesis, the edit distance (number of substitutions, deletions, and insertions) to the most similar reference is calculated.

SSER (subjective sentence error rate): Each translated sentence is judged by a human examiner according to an error scale from 0.0 (semantically and syntactically correct) to 1.0 (completely wrong).

ISER (information item semantic error rate): The test sentences are segmented into information items; for each of these items, the translation candidates are assigned either "OK" or an error class. If the intended information is conveyed, the translation of an information item is considered correct, even if there are slight syntactic errors which do not seriously deteriorate the intelligibility.

For evaluating the SSER and the ISER, we have used the evaluation tool EvalTrans (Nießen and Leusch 2000), which is designed to facilitate the work of manually judging evaluation quality and to ensure consistency over time and across evaluators.

7.4 Impact of the Corpus Size

It is a costly and time-consuming task to compile large texts and have them translated to form bilingual corpora suitable for training the model parameters for statistical machine translation. As a consequence, it is important to investigate the amount of data necessary to sufficiently cover the vocabulary expected in testing. Furthermore, we want to examine to what extent the incorporation of morphological knowledge sources can reduce this amount of necessary data. Figure 4 shows the relation between the size of a typical German corpus and the corresponding number of different full forms. At the size of 520,000 words, the size of the Verbmobil corpus used for training, this curve still has a high growth rate.

To investigate the impact of the size of the bilingual corpus available for training, on translation quality three different setups for training the statistical lexicon on Verbmobil data have been defined:

- using the full training corpus as described in Table 5, comprising 58,000 sentences
- restricting the corpus to 5,000 sentences (approximately every 11th sentence)
- using no bilingual training corpus at all (only a bilingual dictionary; see subsequent discussion)

The language model is always trained on the full English corpus. The argument for this is that monolingual corpora are always easier and less expensive to obtain than bilingual corpora. A conventional dictionary is used in all three setups to complement

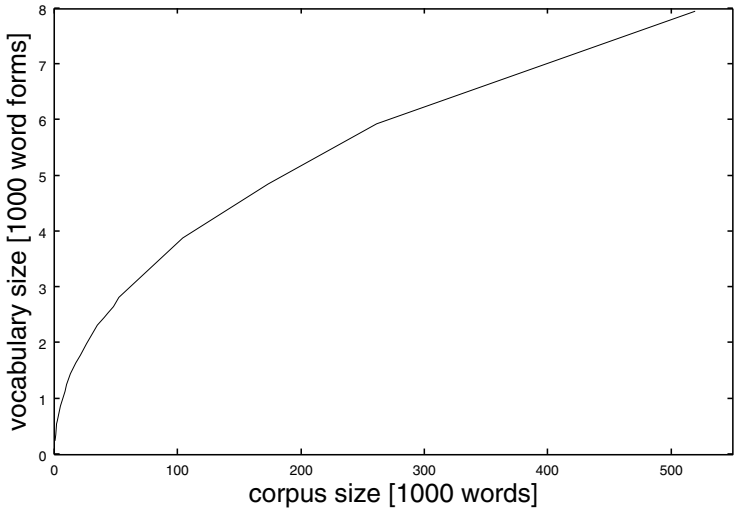

Figure 4
Impact of corpus size (measured in number of running words in the corpus) on vocabulary size (measured in number of different full-form words found in the corpus) for the German part of the Verbmobil corpus.

the bilingual corpus. In the last setup, the lexicon probabilities are trained exclusively on this dictionary

As Table 9 shows, the quality of translation drops significantly when the amount of bilingual data available during training is reduced: When the training corpus is restricted to 5,000 sentences, the SSER increases by about 7% and the ISER by about 3%. As could be expected, the translations produced by the system trained exclusively on a conventional dictionary are very poor: The SSER jumps over 60%.

7.5 Results for Log-Linear Lexicon Combination

7.5.1 Results on the Verbmobil Task. As was pointed out in Section 4, the hierarchical lexicon is expected to be especially useful in cases in which many of the inflected word forms to be accounted for in test do not occur during training. To systematically investigate the model's generalization capability, it has been applied on the three different setups described in Section 7.4. The training procedure was the one proposed in Section 6, which includes restructuring transformations in training and test. Table 9 summarizes the improvement achieved for all three setups.

Training on 58,000 sentences plus conventional dictionary: Compared to the effect of restructuring, the additional improvement achieved with the hierarchical lexicon is relatively small in this setup. The combination of all methods results in a relative improvement in terms of SSER of almost 13% and in terms of information ISER of more than 16% as compared to the baseline.

Training on 5,000 sentences plus conventional dictionary: Restructuring alone can improve the translation quality from 37.3% to 33.6%. The benefit from the hierarchical lexicon is larger in this setup, and the resulting in SSER is 31.8%. This is a relative improvement of almost 15%. The relative improvement in terms of ISER is almost 22%. Note that by applying the methods

Table 9
Results for hierarchical lexicon models and translation with scarce resources. "Restructuring" entails treatment of question inversion and separated verb prefixes as well as merging of phrases in both languages. A conventional dictionary is available in all three setups. The language model is always trained on the full monolingual English corpus. Task: Verbmobil. Testing on 527 sentences (`Test` and `Develop`).

Number of sentences for training		BLEU	m-WER	SSER	ISER
58,000	Baseline	53.7%	34.1%	30.2%	14.1%
	Restructuring	56.3	32.5	26.6	12.8
	+ dictionary disambiguated				
	+ hierarchical lexicon	57.1	31.8	26.3	11.8
5,000	Baseline	47.4	38.0	37.3	17.4
	Restructuring	52.1	34.7	33.6	15.2
	+ dictionary disambiguated				
	+ hierarchical lexicon	52.9	33.9	31.8	13.7
0	Baseline	23.3	53.6	60.4	29.8
	Restructuring	29.1	50.2	57.8	30.0
	+ dictionary disambiguated				
	+ hierarchical lexicon	32.6	48.0	52.8	24.1

proposed here, the corpus for training can be reduced to less than 10% of the original size while increasing the SSER only from 30.2% to 31.8% compared to the baseline when using the full corpus.

Training only on conventional dictionary: In this setup the impact of the hierarchical lexicon is clearly larger than the effect of the restructuring methods, because here the data sparseness problem is much more important than the word order problem. The overall relative reduction in terms of SSER is 13.7% and in terms of ISER 19.1%. An error rate of about 52% is still very poor, but it is close to what might be acceptable when only the gist of the translated document is needed, as is the case in the framework of document classification or multilingual information retrieval.

Examples taken from the Verbmobil Eval-2000 test set are given in Table 10. Smoothing the lexicon probabilities over the inflected forms of the same lemma enables the translation of *sind* as *would* instead of *are*. The smoothed lexicon contains the translation *convenient* for any inflected form of *bequem*. The comparative *more convenient* would be the completely correct translation. The last two examples in the table demonstrate the effect of the disambiguating analyzer, which on the basis of the sentence context identifies *Zimmer* as plural (it has been translated into the singular form *room* by the baseline system) and *das* as an article to be translated by *the* instead of a pronoun which would be translated as *that*. The last example demonstrates that overfitting on domain-specific training can be problematic in some cases: Generally, *because* is a good translation for the co-ordinating conjunction *denn*, but in the appointment-scheduling domain, *denn* is often an adverb, and it often occurs in the same sentence as *dann*, as in *Wie wäre es denn dann?*. The translation for this sentence is something like *How about then?*. Because of the frequency of this domain-specific language use, the word form *denn* is often aligned to *then* in the training corpus. The hierarchical

Table 10
Examples of the effect of the hierarchical lexicon.

Input	sind Sie mit einem Doppelzimmer einverstanden?
Baseline	are you agree with a double room?
Hierarchical lexicon	would you agree with a double room?
Input	mit dem Zug ist es bequemer.
Baseline	by train it is UNKNOWN-bequemer.
Hierarchical lexicon	by train it is convenient.
Input	wir haben zwei Zimmer.
Baseline	we have two room.
Hierarchical lexicon	we have two rooms.
Input	ich würde das Hilton vorschlagen denn es ist das beste.
Baseline	I would suggest that Hilton then it is the best.
Hierarchical lexicon	I would suggest the Hilton because it is the best.

lexicon distinguishes the adverb reading and the conjunction reading, and the correct translation *because* is the highest-ranking one for the conjunction.

7.5.2 Results on the Nespole! Task. We were provided with a small German-English corpus from the Nespole! project (see Section 7.1 for a description). From Table 5 it is obvious that this task is an example of very scarce training data, and it is thus interesting to test the performance of the methods proposed in this article on this task. The same conventional dictionary as was used for the experiments on Verbmobil data (cf. Table 6) complemented the small bilingual training corpus. Furthermore, the (monolingual) English part of the Verbmobil corpus was used in addition to the English part of the Nespole! corpus for training the language model. Table 11 summarizes the results. Information items have not been defined for this test set. An overall relative improvement of 16.5% in the SSER can be achieved.

8. Conclusion

In this article we have proposed methods of incorporating morphological and syntactic information into systems for statistical machine translation. The overall goal was to improve translation quality and to reduce the amount of parallel text necessary to

Table 11
Results for hierarchical lexicon model Nespole! "Restructuring" entails treatment of question inversion and separated verb prefixes as well as merging of phrases in both languages. The same conventional dictionary was used as in the experiments the Verbmobil. The language model was trained on a combination of the English parts of the Nespole! corpus and the Verbmobil corpus.

	BLEU	m-WER	SSER
Baseline	31.6%	50.2%	41.1%
Restructuring	33.7	45.9	38.1
+ hierarchical lexicon	36.5	44.1	34.3

train the model parameters. Substantial improvements on the Verbmobil task and the Nespole! task were achieved.

Some sentence-level restructuring transformations have been introduced which are motivated by knowledge about the sentence structure in the languages involved. These transformations aim at the assimilation of word orders in related sentences.

A hierarchy of equivalence classes has been defined on the basis of morphological and syntactic information beyond the surface forms. The study of the effect of using information from either degree of abstraction led to the construction of hierarchical lexicon models, which combine different items of information in a log-linear way. The benefit from these combined models is twofold: First, the lexical coverage is improved, because the translation of unseen word forms can be derived by considering information from lower levels in the hierarchy. Second, category ambiguity can be resolved, because syntactical context information is made locally accessible by means of annotation with morpho-syntactic tags. As a side effect of the preparative work for setting up the underlying hierarchy of morpho-syntactic information, those pieces of information inherent in fully inflected word forms that are not relevant for translation are detected.

A method for aligning corresponding readings in conventional dictionaries containing pairs of fully inflected word forms has been proposed. The approach uses information deduced from one language side to resolve category ambiguity in the corresponding entry in the other language. The resulting disambiguated dictionaries have proven to be better suited for improving the quality of machine translation, especially if they are used in combination with the hierarchical lexicon models.

The amount of bilingual training data required to achieve an acceptable quality of machine translation has been systematically investigated. All the methods mentioned previously contribute to a better exploitation of the available bilingual data and thus to improving translation quality in frameworks with scarce resources. Three setups for training the parameters of the statistical lexicon on Verbmobil data have been examined: (1) Using the full 58,000 sentences comprising the bilingual training corpus, (2) restricting the corpus to 5,000 sentences, and (3) using only a conventional dictionary. For each of these setups, a relative improvement in terms of subjective sentence error rate between 13% and 15% as compared to the baseline could be obtained using combinations of the methods described in this article. The amount of bilingual training data could be reduced to less than 10% of the original corpus, while losing only 1.6% in accuracy as measured by the subjective sentence error rate. A relative improvement of 16.5% in terms of subjective sentence error rate could also be achieved on the Nespole! task.

Acknowledgments

This work has been partially supported as part of the Verbmobil project (contract number 01 IV 701 T4) by the German Federal Ministry of Education, Science, Research and Technology and as part of the EuTrans project (project number 30268) by the European Union. For the provision of the Nespole! data we thank the Nespole! consortium, listed on the project's home page (Nespole! 2000). Special thanks to Alon Lavie, Lori Levin, Stephan Vogel, and Alex Waibel (in alphabetical order).

References

Al-Onaizan, Yaser, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Kenji Yamada. 2000. Translating with scarce resources. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 672–678, Austin, TX, August.

Berger, Adam L., Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, J. R. Gillett, and A. S. Kehler. 1996. Language translation apparatus and method of using context-based translation models. United States Patent, Patent Number 5510981, April.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and M. J. Goldsmith. 1993. But dictionaries are data too. In *Proceedings of the ARPA Human Language Technology Workshop '93*, pages 202–205, Princeton, NJ, March.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, and Paul S. Roossin. 1988. A statistical approach to language translation. In *Proceedings of COLING 1988: The 12th International Conference on Computational Linguistics*, pages 71–76, Budapest, August.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, John D. Lafferty, and Robert L. Mercer. 1992. Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of TMI 1992: Fourth International Conference on Theoretical and Methodological Issues in MT*, pages 83–100, Montreal, Quebec, Canada, June.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. Mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Darroch, J. N. and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480.

Della Pietra, Stephen A., Vincent J. Della Pietra, and John D. Lafferty. 1995. Inducing features of random fields. Technical Report CMU-CS-95-144, Carnegie Mellon University, Pittsburgh, PA.

Foster, George. 2000. A maximum entropy/minimum divergence translation model. In *Proceedings of ACL 2000: The 38th Annual Meeting of the Association for Computational Linguistics*, pages 37–44, Hong Kong, October.

García-Varea, Ismael and Francisco Casacuberta. 2001. Search algorithms for statistical machine translation based on dynamic programming and pruning techniques. In *Proceedings of the MT Summit VIII*, pages 115–120, Santiago de Compostela, Spain, September.

Germann, Ulrich, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proceedings of ACL-EACL 2001: The 39th Annual Meeting of the Association for Computational Linguistics* (joint with EACL 2001), pages 228–235, Toulouse, France, July.

Kanevsky, Dimitri, Salim Roukos, and Jan Sedivy. 1997. Statistical language model for inflected languages. United States Patent, Patent Number 5835888.

Karlsson, Fred. 1990. Constraint grammar as a framework for parsing running text. In *Proceedings of COLING 1990: The 13th International Conference on Computational Linguistics*, volume 3, pages 168–173, Helsinki, August.

Koehn, Philipp and Kevin Knight. 2001. Knowledge sources for word-level translation models. In Lillian Lee and Donna Harman, editors, *Proceedings of EMNLP 2001: Conference on Empirical Methods in Natural Language Processing*, pages 27–35, Pittsburgh, PA, June.

Larson, Martha, Daniel Willett, Joachim Köhler, and Gerhard Rigoll. 2000. Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches. In *Proceedings ICSLP 2000: Sixth International Conference on Spoken Language Processing*, volume 3, pages 945–948, Beijing, February.

Lavie, Alon, Chad Langley, Alex Waibel, Fabio Pianesi, Gianni Lazzari, Paolo Coletti, Loredana Taddei, and Franco Balducci. 2001. Architecture and design considerations in NESPOLE! A speech translation system for e-commerce applications. In James Allan, editor, *Proceedings of HLT 2001: First International Conference on Human Language Technology Research*, pages 31–39, San Diego, March.

Maltese, G., and F. Mancini. 1992. An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. In *Proceedings of ICASSP 1992: International Conference on Acoustics, Speech and Signal Processing*, pages 157–160, San Francisco, March.

NESPOLE! (NEgotiating through SPOken Language in e-commerce). 2000 Project homepage. Available at http://nespole.itc.it/.

Ney, Hermann, Sonja Nießen, Franz Josef Och, Hassan Sawaf, Christoph Tillmann, and Stephan Vogel. 2000. Algorithms for statistical translation of spoken language. *IEEE Transactions on Speech and Audio Processing*, 8(1):24–36.

Nießen, Sonja and Gregor Leusch. 2000. EvalTrans, a tool for semi-automatic evaluation of machine translation. In *Proceedings of LREC 2000,* Athens. Tool is available at http://www-i6.Informatik.RWTH-Aachen.DE/~niessen/Evaluation/.

Nießen, Sonja and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *Proceedings of COLING 2000: The 18th International Conference on Computational Linguistics*, pages 1081–1085, Saarbrücken, Germany, July.

Nießen, Sonja and Hermann Ney. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of MT Summit VIII*, pages 247–252, Santiago de Compostela, Spain, September.

Nießen, Sonja, Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1998. A DP based search algorithm for statistical machine translation. In *Proceedings of COLING-ACL 1998: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 960–967, Montreal, Quebec, Canada, August.

Och, Franz Josef. 2002. *Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, Computer Science Department, RWTH–University of Technology, Aachen, Germany.

Och, Franz Josef, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of EMNLP 1999: Conference on Empirical Methods in Natural Language Processing*, pages 20–28, University of Maryland, College Park, June.

Och, Franz Josef and Hans Weber. 1998. Improving statistical natural language translation with categories and rules. In *Proceedings of COLING-ACL 1998: The 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 985–989, Montreal, Quebec, Canada, August.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Yorktown Heights, NY, September.

Ratnaparkhi, Adwait. 1997. A simple introduction to maximum entropy models for natural language processing. Technical Report 97–08, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, May.

Tillmann, Christoph and Hermann Ney. 2000. Word re-ordering and DP-based search in statistical machine translation. In *Proceedings of COLING 2000: The 18th International Conference on Computational Linguistics*, pages 850–856, Saarbrücken, Germany, August.