Models of Translational Equivalence among Words

I. Dan Melamed*
West Group

Parallel texts (bitexts) have properties that distinguish them from other kinds of parallel data. First, most words translate to only one other word. Second, bitext correspondence is typically only partial—many words in each text have no clear equivalent in the other text. This article presents methods for biasing statistical translation models to reflect these properties. Evaluation with respect to independent human judgments has confirmed that translation models biased in this fashion are significantly more accurate than a baseline knowledge-free model. This article also shows how a statistical translation model can take advantage of preexisting knowledge that might be available about particular language pairs. Even the simplest kinds of language-specific knowledge, such as the distinction between content words and function words, are shown to reliably boost translation model performance on some tasks. Statistical models that reflect knowledge about the model domain combine the best of both the rationalist and empiricist paradigms.

1. Introduction

The idea of a computer system for translating from one language to another is almost as old as the idea of computer systems. Warren Weaver wrote about mechanical translation as early as 1949. More recently, Brown et al. (1988) suggested that it may be possible to construct machine translation systems automatically. Instead of codifying the human translation process from introspection, Brown and his colleagues proposed machine learning techniques to induce models of the process from examples of its input and output. The proposal generated much excitement, because it held the promise of automating a task that forty years of research have proven very labor-intensive and error-prone. Yet very few other researchers have taken up the cause, partly because Brown et al.'s (1988) approach was quite a departure from the paradigm in vogue at the time.

Formally, Brown et al. (1988) built statistical models of **translational equivalence** (or **translation models**[1], for short). In the context of computational linguistics, translational equivalence is a relation that holds between two expressions with the same meaning, where the two expressions are in different languages. Empirical estimation of statistical translation models is typically based on **parallel texts** or **bitexts**—pairs of texts that are translations of each other. As with all statistical models, the best translation models are those whose parameters correspond best with the sources of variance in the data. Probabilistic translation models whose parameters reflect universal properties of translational equivalence and/or existing knowledge about particular

* D1-66F, 610 Opperman Drive, Eagan, MN 55123. E-mail: dan.melamed@westgroup.com

1 The term translation model, which is standard in the literature, refers to a mathematical relationship between two data sets. In this context, the term implies nothing about the *process* of translation between natural languages, automated or otherwise.

languages and language pairs benefit from the best of both the empiricist and rationalist traditions.

This article presents three such models, along with methods for efficiently estimating their parameters. Each new method is designed to account for an additional universal property of translational equivalence in bitexts:

1. Most word tokens translate to only one word token. I approximate this tendency with a one-to-one assumption.
2. Most text segments are not translated word-for-word. I build an explicit noise model.
3. Different linguistic objects have statistically different behavior in translation. I show a way to condition translation models on different word classes to help account for the variety.

Quantitative evaluation with respect to independent human judgments has shown that each of these three estimation biases significantly improves translation model accuracy over a baseline knowledge-free model. However, these biases will not produce the best possible translation models by themselves. Anyone attempting to build an optimal translation model should infuse it with all available knowledge sources, including syntactic, dictionary, and cognate information. My goal here is only to demonstrate the value of some previously unused kinds of information that are always available for translation modeling, and to show how these information sources can be integrated with others.

A review of some previously published translation models follows an introduction to translation model taxonomy. The core of the article is a presentation of the model estimation biases described above. The last section reports the results of experiments designed to evaluate these innovations.

Throughout this article, I shall use $\mathcal{CALLIGRAPHIC}$ letters to denote entire text corpora and other sets of sets, CAPITAL letters to denote collections, including sequences and bags, and *italics* for scalar variables. I shall also distinguish between **types** and tokens by using **bold font** for the former and plain font for the latter.

2. Translation Model Decomposition

There are two kinds of applications of translation models: those where word order plays a crucial role and those where it doesn't. Empirically estimated models of translational equivalence among word types can play a central role in both kinds of applications.

Applications where word order is not essential include

- cross-language information retrieval (e.g., McCarley 1999),
- multilingual document filtering (e.g., Oard 1997),
- computer-assisted language learning (e.g., Nerbonne et al. 1997),
- certain machine-assisted translation tools (e.g., Macklovitch 1994; Melamed 1996a),
- concordancing for bilingual lexicography (e.g., Catizone, Russell, and Warwick 1989; Gale and Church 1991),

- corpus linguistics (e.g., Svartvik 1992),
- "crummy" machine translation (e.g., Church and Hovy 1992; Resnik 1997).

For these applications, empirically estimated models have a number of advantages over handcrafted models such as on-line versions of bilingual dictionaries. Two of the advantages are the possibility of better coverage and the possibility of frequent updates by nonexpert users to keep up with rapidly evolving vocabularies.

A third advantage is that statistical models can provide more accurate information about the relative importance of different translations. Such information is crucial for applications such as cross-language information retrieval (CLIR). In the vector space approach to CLIR, the query vector Q' is in a different language (a different vector space) from the document vectors D. A word-to-word translation model T can map Q' into a vector Q in the vector space of D. In order for the mapping to be accurate, T must be able to encode many levels of relative importance among the possible translations of each element of Q'. A typical bilingual dictionary says only what the possible translations are, which is equivalent to positing a uniform translational distribution. The performance of cross-language information retrieval with a uniform T is likely to be limited in the same way as the performance of conventional information retrieval without term-frequency information, i.e., where the system knows which terms occur in which documents, but not how often (Buckley 1993).

Applications where word order is crucial include speech transcription for translation (Brousseau et al. 1995), bootstrapping of OCR systems for new languages (Philip Resnik and Tapas Kanungo, personal communication), interactive translation (Foster, Isabelle, and Plamondon 1996), and fully automatic high-quality machine translation (e.g., Al-Onaizan et al. 1999). In such applications, a word-to-word translation model can serve as an independent module in a more complex sequence-to-sequence translation model.[2] The independence of such a module is desirable for two reasons, one practical and one philosophical. The practical reason is illustrated in this article: Order-independent translation models can be accurately estimated more efficiently in isolation. The philosophical reason is that words are an important epistemological category in our naive mental representations of language. We have many intuitions (and even some testable theories) about what words are and how they behave. We can bring these intuitions to bear on our translation models without being distracted by other facets of language, such as phrase structure. For example, the translation models presented in the last two chapters of Melamed (to appear) capture the intuitions that words can have multiple senses and that spaces in text do not necessarily delimit words.

The independence of a word-to-word translation module in a sequence-to-sequence translation model can be effected by a two-stage decomposition. The first stage is based on the observation that every sequence $\mathbf{L}$ is just an ordered bag, and that the bag $\mathbf{B}$ can be modeled independently of its order $\mathbf{O}$. For example, the sequence $\langle abc \rangle$ consists of the bag $\{c, a, b\}$ and the ordering relation $\{(b, 2), (a, 1), (c, 3)\}$. If we represent each sequence $\mathbf{L}$ as a pair **(B, O)**, then

$$\Pr(\mathbf{L}) \equiv \Pr(\mathbf{B}, \mathbf{O}) \tag{1}$$

$$= \Pr(\mathbf{B}) \cdot \Pr(\mathbf{O}|\mathbf{B}). \tag{2}$$

2 "Sentence-to-sentence" might be a more transparent term than "sequence-to-sequence," but all the models that I'm aware of apply equally well to sequences of words that are not sentences.

Now, let $\mathbf{L_1}$ and $\mathbf{L_2}$ be two sequences and let $\mathbf{A}$ be a one-to-one mapping between the elements of $\mathbf{L_1}$ and the elements of $\mathbf{L_2}$. Borrowing a term from the operations research literature, I shall refer to such mappings as **assignments**.[3] Let $\mathcal{A}$ be the set of all possible assignments between $\mathbf{L_1}$ and $\mathbf{L_2}$. Using assignments, we can decompose conditional and joint probabilities over sequences:

$$\Pr(\mathbf{L_1}|\mathbf{L_2}) = \sum_{A \in \mathcal{A}} \Pr(\mathbf{L_1}, \mathbf{A}|\mathbf{L_2}) \tag{3}$$

$$\Pr(\mathbf{L_1}, \mathbf{L_2}) = \sum_{A \in \mathcal{A}} \Pr(\mathbf{L_1}, \mathbf{A}, \mathbf{L_2}) \tag{4}$$

where

$$\Pr(\mathbf{L_1}, \mathbf{A}|\mathbf{L_2}) \equiv \Pr(\mathbf{B_1}, \mathbf{O_1}, \mathbf{A}|\mathbf{L_2}) \tag{5}$$

$$= \Pr(\mathbf{B_1}, \mathbf{A}|\mathbf{L_2}) \cdot \Pr(\mathbf{O_1}|\mathbf{B_1}, \mathbf{A}, \mathbf{L_2}) \tag{6}$$

$$\Pr(\mathbf{L_1}, \mathbf{A}, \mathbf{L_2}) \equiv \Pr(\mathbf{B_1}, \mathbf{O_1}, \mathbf{A}, \mathbf{B_2}, \mathbf{O_2}) \tag{7}$$

$$= \Pr(\mathbf{B_1}, \mathbf{A}, \mathbf{B_2}) \cdot \Pr(\mathbf{O_1}, \mathbf{O_2}|\mathbf{B_1}, \mathbf{A}, \mathbf{B_2}) \tag{8}$$

Summing bag pair probabilities over all possible assignments, we obtain a **bag-to-bag translation model**:

$$\Pr(\mathbf{B_1}, \mathbf{B_2}) = \sum_{A \in \mathcal{A}} \Pr(\mathbf{B_1}, \mathbf{A}, \mathbf{B_2}) \tag{9}$$

The second stage of decomposition takes us from bags of words to the words that they contain. The following bag pair generation process illustrates how a word-to-word translation model can be embedded in a bag-to-bag translation model for languages $\mathcal{L}_1$ and $\mathcal{L}_2$:

1. Generate a bag size l.[4] l is also the assignment size.
2. Generate l language-independent concepts $C_1, \ldots, C_l$.
3. From each concept C_i, $1 \leq i \leq l$, generate a pair of word sequences $(\vec{u}_i, \vec{v}_i)$ from $\mathcal{L}_1^* \times \mathcal{L}_2^*$, according to the distribution $trans(\vec{\mathbf{u}}, \vec{\mathbf{v}})$, to lexicalize the concept in the two languages.[5] Some concepts are not lexicalized in some languages, so one of $\vec{u}_i$ and $\vec{v}_i$ may be empty.

A pair of bags containing m and n nonempty word sequences can be generated by a process where l is anywhere between 1 and $m + n$.

For notational convenience, the elements of the two bags can be labeled so that $\mathbf{B_1} \equiv \{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_l\}$ and $\mathbf{B_2} \equiv \{\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_l\}$, where some of the $\vec{\mathbf{u}}$'s and $\vec{\mathbf{v}}$'s may be empty. The elements of an assignment, then, are pairs of bag element labels: $\mathbf{A} \equiv \{(i_1, j_1), \ldots, (i_l, j_l)\}$, where each i ranges over $\{\vec{\mathbf{u}}_1, \ldots, \vec{\mathbf{u}}_l\}$, each j ranges over $\{\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_l\}$,

3 Assignments are different from Brown, Della Pietra, Della Pietra, and Mercer's (1993) alignments in that assignments can range over pairs of arbitrary labels, not necessarily sequence position indexes. Also, unlike alignments, assignments must be one-to-one.

4 The exact nature of the bag size distribution is immaterial for the present purposes.

5 Since they are put into bags, $\vec{u}_i$ and $\vec{v}_i$ could just as well be bags instead of sequences. I make them sequences only to be consistent with more sophisticated models that account for noncompositional compounds (e.g. Melamed, to appear, Chapter 8).

each i is distinct, and each j is distinct. The label pairs in a given assignment can be generated in any order, so there are $l!$ ways to generate an assignment of size l.[6] It follows that the probability of generating a pair of bags $(\mathbf{B_1}, \mathbf{B_2})$ with a particular assignment $\mathbf{A}$ of size l is

$$\Pr(\mathbf{B}_1, \mathbf{A}, \mathbf{B}_2 | l, C, trans) = \Pr(l) \cdot l! \prod_{(i,j) \in \mathbf{A}} \sum_{C \in \mathcal{C}} \Pr(C) trans(\vec{\mathbf{u}}_i, \vec{\mathbf{v}}_i | C). \quad (10)$$

The above equation holds regardless of how we represent concepts. There are many plausible representations, such as pairs of trees from synchronous tree adjoining grammars (Abeillé et al. 1990; Shieber 1994; Candito 1998), lexical conceptual structures (Dorr 1992) and WordNet synsets (Fellbaum 1998; Vossen 1998). Of course, for a representation to be used, a method must exist for estimating its distribution in data. A useful representation will reduce the entropy of the *trans* distribution, which is conditioned on the concept distribution as shown in Equation 10. This topic is beyond the scope of this article, however. I mention it only to show how the models presented here may be used as building blocks for models that are more psycholinguistically sophisticated.

To make the translation model estimation methods presented here as general as possible, I shall assume a totally uninformative concept representation—the *trans* distribution itself. In other words, I shall assume that each different pair of word sequence types is deterministically generated from a different concept, so that $trans(\vec{\mathbf{u}}_i, \vec{\mathbf{v}}_i | C)$ is zero for all concepts except one. Now, a bag-to-bag translation model can be fully specified by the distributions of l and *trans*.

$$\Pr(\mathbf{B}_1, \mathbf{A}, \mathbf{B}_2 | l, trans) = \Pr(l) \cdot l! \prod_{(i,j) \in \mathbf{A}} trans(\vec{\mathbf{u}}_i, \vec{\mathbf{v}}_j) \quad (11)$$

The probability distribution $trans(\vec{\mathbf{u}}, \vec{\mathbf{v}})$ is a word-to-word translation model. Unlike the models proposed by Brown et al. (1993b), this model is **symmetric**, because both word bags are generated together from a joint probability distribution. Brown and his colleagues' models, reviewed in Section 4.3, generate one half of the bitext given the other half, so they are represented by conditional probability distributions. A sequence-to-sequence translation model can be obtained from a word-to-word translation model by combining Equation 11 with order information as in Equation 8.

3. The One-to-One Assumption

The most general word-to-word translation model $trans(\vec{\mathbf{u}}, \vec{\mathbf{v}})$, where $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ range over sequences in $\mathcal{L}_1$ and $\mathcal{L}_2$, has an infinite number of parameters. This model can be constrained in various ways to make it more practical. The models presented in this article are based on the **one-to-one assumption**: Each word is translated to at most one other word. In these models, $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$ may consist of at most one word each. As before, one of the two sequences (but not both) may be empty. I shall describe empty sequences as consisting of a special NULL word, so that each word sequence will contain exactly one word and can be treated as a scalar. Henceforth, I shall write $\mathbf{u}$ and $\mathbf{v}$ instead of $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$. Under the one-to-one assumption, a pair of bags containing m

6 The number of permutations is smaller when either bag contains two or more identical elements, but this detail will not affect the estimation algorithms presented here.

and n nonempty words can be generated by a process where the bag size l is anywhere between $\max(m, n)$ and $m + n$.

The one-to-one assumption is not as restrictive as it may appear: The explanatory power of a model based on this assumption may be raised to an arbitrary level by extending Western notions of what words are to include words that contain spaces (e.g., in English) or several characters (e.g., in Chinese). For example, I have shown elsewhere how to estimate word-to-word translation models where a word can be a noncompositional compound consisting of several space-delimited tokens (Melamed, to appear). For the purposes of this article, however, **words** are the tokens generated by my tokenizers and stemmers for the languages in question. Therefore, the models in this article are only a first approximation to the vast complexities of translational equivalence between natural languages. They are intended mainly as stepping stones towards better models.

4. Previous Work

4.1 Models of Co-occurrence

Most methods for estimating translation models from bitexts start with the following intuition: Words that are translations of each other are more likely to appear in corresponding bitext regions than other pairs of words. Following this intuition, most authors begin by counting the number of times that word types in one half of the bitext co-occur with word types in the other half. Different co-occurrence counting methods stem from different models of co-occurrence.

A model of co-occurrence is a Boolean predicate, which indicates whether a given pair of word *tokens* co-occur in corresponding regions of the bitext space. Different models of co-occurrence are possible, depending on the kind of bitext map that is available, the language-specific information that is available, and the assumptions made about the nature of translational equivalence. All the translation models reviewed and introduced in this article can be based on any of the co-occurrence models described by Melamed (1998a). For expository purposes, however, I shall assume a boundary-based model of co-occurrence throughout this article. A boundary-based model of co-occurrence assumes that both halves of the bitext have been segmented into s segments, so that segment U_i in one half of the bitext and segment V_i in the other half are mutual translations, $1 \leq i \leq s$.

Under the boundary-based model of co-occurrence, there are several ways to compute co-occurrence counts $cooc(\mathbf{u}, \mathbf{v})$ between word types $\mathbf{u}$ and $\mathbf{v}$. In the models of Brown, Della Pietra, Della Pietra, and Mercer (1993), reviewed in Section 4.3,

$$cooc(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{s} e_i(\mathbf{u}) \cdot f_i(\mathbf{v}), \tag{12}$$

where e_i and f_i are the unigram frequencies of $\mathbf{u}$ and $\mathbf{v}$, respectively, in each aligned text segment i. For most translation models, this method produces suboptimal results, however, when $e_i(\mathbf{u}) > 1$ and $f_i(\mathbf{v}) > 1$. I argue elsewhere (Melamed 1998a) that

$$cooc(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{s} \min[e_i(\mathbf{u}), f_i(\mathbf{v})] \tag{13}$$

is preferable, and this is the method used for the models introduced in Section 5.

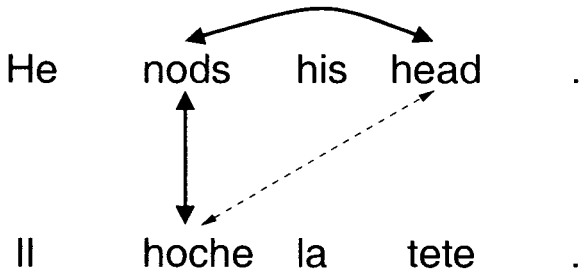


Figure 1
nods and *hoche* often co-occur, as do *nods* and *head*. The direct association between *nods* and *hoche*, and the direct association between *nods* and *head* give rise to an indirect association between *hoche* and *head*.

4.2 Nonprobabilistic Translation Lexicons

Many researchers have proposed greedy algorithms for estimating nonprobabilistic word-to-word translation models, also known as translation lexicons (e.g., Catizone, Russell, and Warwick 1989; Gale and Church 1991; Fung 1995; Kumano and Hirakawa 1994; Melamed 1995; Wu and Xia 1994). Most of these algorithms can be summarized as follows:

1. Choose a similarity function S between word types in $\mathcal{L}_1$ and word types in $\mathcal{L}_2$.
2. Compute association scores $S(\mathbf{u}, \mathbf{v})$ for a set of word type pairs $(\mathbf{u}, \mathbf{v}) \in (\mathcal{L}_1 \times \mathcal{L}_2)$ that occur in training data.
3. Sort the word pairs in descending order of their association scores.
4. Discard all word pairs for which $S(\mathbf{u}, \mathbf{v})$ is less than a chosen threshold. The remaining word pairs become the entries in the translation lexicon.

The various proposals differ mainly in their choice of similarity function. Almost all the similarity functions in the literature are based on a model of co-occurrence with some linguistically motivated filtering (see Fung [1995] for a notable exception).

Given a reasonable similarity function, the greedy algorithm works remarkably well, considering how simple it is. However, the association scores in Step 2 are typically computed independently of each other. The problem with this independence assumption is illustrated in Figure 1. The two word sequences represent corresponding regions of an English/French bitext. If *nods* and *hoche* co-occur much more often than expected by chance, then any reasonable similarity metric will deem them likely to be mutual translations. *Nods* and *hoche* are indeed mutual translations, so their tendency to co-occur is called a **direct association**. Now, suppose that *nods* and *head* often co-occur in English. Then *hoche* and *head* will also co-occur more often than expected by chance. The dashed arrow between *hoche* and *head* in Figure 1 represents an **indirect association**, since the association between *hoche* and *head* arises only by virtue of the association between each of them and *nods*. Models of translational equivalence that are ignorant of indirect associations have "a tendency . . . to be confused by collocates" (Dagan, Church, and Gale 1993,5).

Paradoxically, the irregularities (noise) in text and in translation mitigate the problem. If noise in the data reduces the strength of a direct association, then the same noise will reduce the strengths of any indirect associations that are based on this direct

Table 1
Variables used to describe translation models.

$(\mathcal{U}, \mathcal{V})$	=	the two halves of the bitext
(U, V)	=	a pair of aligned text segments in $(\mathcal{U}, \mathcal{V})$
$e(\mathbf{u})$	=	the unigram frequency of $\mathbf{u}$ in U
$f(\mathbf{v})$	=	the unigram frequency of $\mathbf{v}$ in V
$cooc(\mathbf{u}, \mathbf{v})$	=	the number of times that $\mathbf{u}$ and $\mathbf{v}$ co-occur
$trans(\mathbf{v}\|\mathbf{u})$	=	the probability that a token of $\mathbf{u}$ will be translated as a token of $\mathbf{v}$

association. On the other hand, noise can reduce the strength of an indirect association without affecting any direct associations. Therefore, direct associations are usually stronger than indirect associations. If all the entries in a translation lexicon are sorted by their association scores, the direct associations will be very dense near the top of the list, and sparser towards the bottom.

Gale and Church (1991) have shown that entries at the very top of the list can be over 98% correct. Their algorithm gleaned lexicon entries for about 61% of the word tokens in a sample of 800 English sentences. To obtain 98% precision, their algorithm selected only entries for which it had high confidence that the association score was high. These would be the word pairs that co-occur most frequently. A random sample of 800 sentences from the same corpus showed that 61% of the word tokens, where the tokens are of the most frequent types, represent 4.5% of all the word types.

A similar strategy was employed by Wu and Xia (1994) and by Fung (1995). Fung skimmed off the top 23.8% of the noun-noun entries in her lexicon to achieve a precision of 71.6%. Wu and Xia have reported automatic acquisition of 6,517 lexicon entries from a 3.3-million-word corpus, with a precision of 86%. The first 3.3 million word tokens in an English corpus from a similar genre contained 33,490 different word types, suggesting a recall of roughly 19%. Note, however, that Wu and Xia chose to weight their precision estimates by the probabilities attached to each entry:

> For example, if the translation set for English word *detect* has the two correct Chinese candidates with 0.533 probability and with 0.277 probability, and the incorrect translation with 0.190 probability, then we count this as 0.810 correct translations and 0.190 incorrect translations. (Wu and Xia 1994, 211)

This is a reasonable evaluation method, but it is not comparable to methods that simply count each lexicon entry as either right or wrong (e.g., Daille, Gaussier, and Langé 1994; Melamed 1996b). A weighted precision estimate pays more attention to entries that are more frequent and hence easier to estimate. Therefore, weighted precision estimates are generally higher than unweighted ones.

4.3 Reestimated Sequence-to-Sequence Translation Models

Most probabilistic translation model reestimation algorithms published to date are variations on the theme proposed by Brown et al. (1993b). These models involve conditional probabilities, but they can be compared to symmetric models if the latter are normalized by the appropriate marginal distribution. I shall review these models using the notation in Table 1.

4.3.1 Models Using Only Co-occurrence Information. Brown and his colleagues employ the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) to estimate the parameters of their Model 1. On iteration i, the EM algorithm reestimates the model parameters $trans_i(\mathbf{v}|\mathbf{u})$ based on their estimates from iteration $i-1$. In Model 1, the relationship between the new parameter estimates and the old ones is

$$trans_i(\mathbf{v}|\mathbf{u}) = z \sum_{(U,V)\in(\mathcal{U},\mathcal{V})} \frac{trans_{i-1}(\mathbf{v}|\mathbf{u}) \cdot e(\mathbf{u}) \cdot f(\mathbf{v})}{\sum_{\mathbf{u}'\in U} trans_{i-1}(\mathbf{v}|\mathbf{u}')} \tag{14}$$

where z is a normalizing factor.[7]

It is instructive to consider the form of Equation 14 when all the translation probabilities $trans(\mathbf{v}|\mathbf{u})$ for a particular $\mathbf{u}$ are initialized to the same constant p, as Brown et al. (1993b, 273) actually do:

$$trans_1(\mathbf{v}|\mathbf{u}) = z \sum_{(U,V)\in(\mathcal{U},\mathcal{V})} \frac{p \cdot e(\mathbf{u}) \cdot f(\mathbf{v})}{p \cdot |U|} \tag{15}$$

$$= z \sum_{(U,V)\in(\mathcal{U},\mathcal{V})} \frac{e(\mathbf{u}) \cdot f(\mathbf{v})}{|U|} \tag{16}$$

The initial translation probability $trans_1(\mathbf{v}|\mathbf{u})$ is set proportional to the co-occurrence count of $\mathbf{u}$ and $\mathbf{v}$ and inversely proportional to the length of each segment U in which $\mathbf{u}$ occurs. The intuition behind the numerator is central to most bitext-based translation models: The more often two words co-occur, the more likely they are to be mutual translations. The intuition behind the denominator is that the co-occurrence count of $\mathbf{u}$ and $\mathbf{v}$ should be discounted to the degree that $\mathbf{v}$ also co-occurs with other words in the same segment pair.

Now consider how Equation 16 would behave if all the text segments on each side were of the same length,[8] so that each token of $\mathbf{v}$ co-occurs with exactly c words (where c is constant):

$$trans_1(\mathbf{v}|\mathbf{u}) = z \sum_{(U,V)\in(\mathcal{U},\mathcal{V})} \frac{e(\mathbf{u}) \cdot f(\mathbf{v})}{c} \tag{17}$$

$$= \frac{z}{c} \sum_{(U,V)\in(\mathcal{U},\mathcal{V})} e(\mathbf{u}) \cdot f(\mathbf{v}) \tag{18}$$

The normalizing coefficient $\frac{z}{c}$ is constant over all words. The only difference between Equations 16 and 18 is that the former discounts co-occurrences proportionally to the segment lengths. When information about segment lengths is not available, the only information available to initialize Model 1 is the co-occurrence counts. This property makes Model 1 an appropriate baseline for comparison to more sophisticated models that use other information sources, both in the work of Brown and his colleagues and in the work described here.

7 This expression is obtained by substituting Brown, Della Pietra, Della Pietra, and Mercer's (1993) Equation 17 into their Equation 14.

8 Or, equivalently, if the notion of segments were dispensed with altogether, as under the distance-based model of co-occurrence (Melamed 1998a).

4.3.2 Word Order Correlation Biases. In any bitext, the positions of words relative to the true bitext map correlate with the positions of their translations. The correlation is stronger for language pairs with more similar word order. Brown et al. (1988) introduced the idea that this correlation can be encoded in translation model parameters. Dagan, Church, and Gale (1993) expanded on this idea by replacing Brown et al.'s (1988) word alignment parameters, which were based on absolute word positions in aligned segments, with a much smaller set of relative offset parameters. The much smaller number of parameters allowed Dagan, Church, and Gale's model to be effectively trained on much smaller bitexts. Vogel, Ney, and Tillmann (1996) have shown how some additional assumptions can turn this model into a hidden Markov model, enabling even more efficient parameter estimation.

It cannot be overemphasized that the word order correlation bias is just knowledge about the problem domain, which can be used to guide the search for the optimum model parameters. Translational equivalence can be empirically modeled for any pair of languages, but some models and model biases work better for some language pairs than for others. The word order correlation bias is most useful when it has high predictive power, i.e., when the distribution of alignments or offsets has low entropy. The entropy of this distribution is indeed relatively low for the language pair that both Brown and his colleagues and Dagan, Church, and Gale were working with—French and English have very similar word order. A word order correlation bias, as well as the phrase structure biases in Brown et al.'s (1993b) Models 4 and 5, would be less beneficial with noisier training bitexts or for language pairs with less similar word order. Nevertheless, one should use all available information sources, if one wants to build the best possible translation model. Section 5.3 suggests a way to add the word order correlation bias to the models presented in this article.

4.4 Reestimated Bag-to-Bag Translation Models

At about the same time that I developed the models in this article, Hiemstra (1996) independently developed his own bag-to-bag model of translational equivalence. His model is also based on a one-to-one assumption, but it differs from my models in that it allows empty words in only one of the two bags, the one representing the shorter sentence. Thus, Hiemstra's model is similar to the first model in Section 5, but it has a little less explanatory power. Hiemstra's approach also differs from mine in his use of the Iterative Proportional Fitting Procedure (IPFP) (Deming and Stephan 1940) for parameter estimation.

The IPFP is quite sensitive to initial conditions, so Hiemstra investigated a number of initialization options. Choosing the most advantageous, Hiemstra has published parts of the translational distributions of certain words, induced using both his method and Brown et al.'s (1993b) Model 1 from the same training bitext. Subjective comparison of these examples suggests that Hiemstra's method is more accurate. Hiemstra (1998) has also evaluated the recall and precision of his method and of Model 1 on a small hand-constructed set of link tokens in a particular bitext. Model 1 fared worse, on average.

5. Parameter Estimation

This section describes my methods for estimating the parameters of a symmetric word-to-word translation model from a bitext. For most applications, we are interested in estimating the probability $trans(\mathbf{u}, \mathbf{v})$ of jointly generating the pair of words $(\mathbf{u}, \mathbf{v})$. Unfortunately, these parameters cannot be directly inferred from a training bitext, because we don't know which words in one half of the bitext were generated together

with which words in the other half. The observable features of the bitext are only the co-occurrence counts $cooc(\mathbf{u}, \mathbf{v})$ (see Section 4.1).

Methods for estimating translation parameters from co-occurrence counts typically involve **link counts** $links(\mathbf{u}, \mathbf{v})$, which represent hypotheses about the number of times that $\mathbf{u}$ and $\mathbf{v}$ were generated together, for each $\mathbf{u}$ and $\mathbf{v}$ in the bitext. A **link token** is an ordered pair of word tokens, one from each half of the bitext. A **link type** is an ordered pair of word types. The link counts $links(\mathbf{u}, \mathbf{v})$ range over link types. We can always estimate $trans(\mathbf{u}, \mathbf{v})$ by normalizing link counts so that $\sum_{\mathbf{u},\mathbf{v}} trans(\mathbf{u}, \mathbf{v}) = 1$:

$$trans(\mathbf{u}, \mathbf{v}) = \frac{links(\mathbf{u}, \mathbf{v})}{\sum_{\mathbf{u}',\mathbf{v}'} links(\mathbf{u}', \mathbf{v}')} \tag{19}$$

For estimation purposes, it is convenient to also employ a separate set of non-probabilistic parameters $score(\mathbf{u}, \mathbf{v})$, which represent the chances that $\mathbf{u}$ and $\mathbf{v}$ can ever be mutual translations, i.e., that there exists some context where tokens u and v are generated from the same concept. The relationship between $score(\mathbf{u}, \mathbf{v})$ and $trans(\mathbf{u}, \mathbf{v})$ can be more or less direct, depending on the model and its estimation method. Each of the models presented below uses a different *score* formulation.

All my methods for estimating the translation parameters $trans(\mathbf{u}, \mathbf{v})$ share the following general outline:

1. Initialize the *score* parameters to a first approximation, based only on the co-occurrence counts.
2. Approximate the expected link counts $links(\mathbf{u}, \mathbf{v})$, as a function of the *score* parameters and the co-occurrence counts.
3. Estimate $trans(\mathbf{u}, \mathbf{v})$, by normalizing the link counts as in Equation 19. If less than .0001 of the $trans(\mathbf{u}, \mathbf{v})$ distribution changed from the previous iteration, then stop.
4. Reestimate the parameters $score(\mathbf{u}, \mathbf{v})$, as a function of the link counts and the co-occurrence counts.
5. Repeat from Step 2.

Under certain conditions, a parameter estimation process of this sort is an instance of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). As explained below, meeting these conditions is computationally too expensive for my models.[9] Therefore, I employ some approximations, which lack the EM algorithm's convergence guarantee.

The maximum likelihood approach to estimating the unknown parameters is to find the set of parameters $\widehat{\Theta}$ that maximize the probability of the training bitext (U, V).

$$\widehat{\Theta} = \arg\max_{\Theta} \Pr(U, V|\Theta) \tag{20}$$

The probability of the bitext is a sum over the distribution $\mathcal{A}$ of possible assignments:

$$\Pr(U, V|\Theta) = \sum_{A \in \mathcal{A}} \Pr(U, A, V|\Theta). \tag{21}$$

9 For example, the expectation in Step 2 would need to be computed exactly, rather than merely approximated.

The number of possible assignments grows exponentially with the size of aligned text segments in the bitext. Due to the parameter interdependencies introduced by the one-to-one assumption, we are unlikely to find a method for decomposing the assignments into parameters that can be estimated independently of each other as in Brown et al. [1993b, Equation 26]). Barring such a decomposition method, the MLE approach is infeasible. This is why we must make do with approximations to the EM algorithm.

In this situation, Brown et al. (1993b, 293) recommend "evaluating the expectations using only a single, probable alignment." The single most probable assignment A_{max} is the **maximum a posteriori (MAP) assignment**:

$$A_{max} = \arg\max_{A \in \mathcal{A}} \Pr(U, A, V|\Theta) \tag{22}$$

$$= \arg\max_{A \in \mathcal{A}} \Pr(l) \cdot l! \prod_{(i,j) \in A} trans(\mathbf{u}_i, \mathbf{v}_j) \tag{23}$$

$$= \arg\max_{A \in \mathcal{A}} \log \left[\Pr(l) \cdot l! \prod_{(i,j) \in A} trans(\mathbf{u}_i, \mathbf{v}_j) \right] \tag{24}$$

$$= \arg\max_{A \in \mathcal{A}} \left\{ \log[\Pr(l) \cdot l!] + \sum_{(i,j) \in A} \log trans(\mathbf{u}_i, \mathbf{v}_j) \right\} \tag{25}$$

To simplify things further, let us assume that $\Pr(l) \cdot l!$ is constant, so that

$$A_{max} = \arg\max_{A \in \mathcal{A}} \sum_{(i,j) \in A} \log trans(\mathbf{u}_i, \mathbf{v}_j). \tag{26}$$

If we represent the bitext as a bipartite graph and weight the edges by $\log trans(\mathbf{u}, \mathbf{v})$, then the right-hand side of Equation 26 is an instance of the weighted maximum matching problem and A_{max} is its solution. For a bipartite graph $G = (V_1 \cup V_2, E)$, with $v = |V_1 \cup V_2|$ and $e = |E|$, the lowest currently known upper bound on the computational complexity of this problem is $O(ve + v^2 \log v)$ (Ahuja, Magnati, and Orlin 1993, 500). Although this upper bound is polynomial, it is still too expensive for typical bitexts.[10] Subsection 5.1.2 describes a greedy approximation to the MAP approximation.

5.1 Method A: The Competitive Linking Algorithm

5.1.1 Step 1: Initialization. Almost every translation model estimation algorithm exploits the well-known correlation between translation probabilities and co-occurrence counts. Many algorithms also normalize the co-occurrence counts $cooc(\mathbf{u}, \mathbf{v})$ by the marginal frequencies of $\mathbf{u}$ and $\mathbf{v}$. However, these quantities account for only the three shaded cells in Table 2. The statistical interdependence between two word types can be estimated more robustly by considering the whole table. For example, Gale and Church (1991, 154) suggest that "ϕ^2, a χ^2-like statistic, seems to be a particularly good choice because it makes good use of the off-diagonal cells" in the contingency table.

10 At least for my current very inefficient implementation.

Table 2
A co-occurrence contingency table.

	$\mathbf{u}$	$\neg\mathbf{u}$	Total
$\mathbf{v}$	$cooc(\mathbf{u}, \mathbf{v})$	$cooc(\neg\mathbf{u}, \mathbf{v})$	$cooc(\cdot, \mathbf{v})$
$\neg\mathbf{v}$	$cooc(\mathbf{u}, \neg\mathbf{v})$	$cooc(\neg\mathbf{u}, \neg\mathbf{v})$	$cooc(\cdot, \neg\mathbf{v})$
Total	$cooc(\mathbf{u}, \cdot)$	$cooc(\neg\mathbf{u}, \cdot)$	$cooc(\cdot, \cdot)$

In informal experiments described elsewhere (Melamed 1995), I found that the G^2 statistic suggested by Dunning (1993) slightly outperforms ϕ^2. Let the cells of the contingency table be named as follows:

	$\mathbf{u}$	$\neg\mathbf{u}$
$\mathbf{v}$	a	b
$\neg\mathbf{v}$	c	d

Now,

$$G^2(\mathbf{u}, \mathbf{v}) = -2 \log \frac{B(a|a+b, p_1)B(c|c+d, p_2)}{B(a|a+b, p)B(c|c+d, p)} \tag{27}$$

where $B(k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$ are binomial probabilities. The statistic uses maximum likelihood estimates for the probability parameters: $p_1 = \frac{a}{a+b}$, $p_2 = \frac{c}{c+d}$, $p = \frac{a+c}{a+b+c+d}$. G^2 is easy to compute because the binomial coefficients in the numerator and in the denominator cancel each other out. All my methods initialize the parameters $score(\mathbf{u}, \mathbf{v})$ to $G^2(\mathbf{u}, \mathbf{v})$, except that any pairing with NULL is initialized to an infinitesimal value. I have also found it useful to smooth the co-occurrence counts, e.g., using the Simple Good-Turing smoothing method (Gale and Sampson 1995), before computing G^2.

5.1.2 Step 2: Estimation of Link Counts. To further reduce the complexity of estimating link counts, I employ the **competitive linking algorithm**, which is a greedy approximation to the MAP approximation:

1. Sort all the $score(\mathbf{u}, \mathbf{v})$ from highest to lowest.
2. For each $score(\mathbf{u}, \mathbf{v})$, in order:
   (a) If $\mathbf{u}$ (resp., $\mathbf{v}$) is NULL, consider all tokens of $\mathbf{v}$ (resp., $\mathbf{u}$) in the bitext linked to NULL. Otherwise, link all co-occurring token pairs (u, v) in the bitext.
   (b) The one-to-one assumption implies that linked words cannot be linked again. Therefore, remove all linked word tokens from their respective halves of the bitext.

The competitive linking algorithm can be viewed as a heuristic search for the most likely assignment in the space of all possible assignments. The heuristic is that the most likely assignments contain links that are individually the most likely. The search proceeds by a process of elimination. In the first search iteration, all the assignments that do not contain the most likely link are discarded. In the second iteration, all the assignments that do not contain the second most likely link are discarded, and

so on until only one assignment remains.[11] The algorithm greedily selects the most likely links first, and then selects less likely links only if they don't conflict with previous selections. The probability of a link being rejected increases with the number of links that are selected before it, and thus decreases with the link's *score*. In this problem domain, the competitive linking algorithm usually finds one of the most likely assignments, as I will show in Section 6. Under an appropriate hashing scheme, the expected running time of the competitive linking algorithm is linear in the size of the input bitext.

The competitive linking algorithm and its one-to-one assumption are potent weapons against the ever-present sparse data problem. They enable accurate estimation of translational distributions even for words that occur only once, as long as the surrounding words are more frequent. In most translation models, link scores are correlated with co-occurrence frequency. So, links between tokens u and v for which $score(\mathbf{u}, \mathbf{v})$ is highest are the ones for which there is the most evidence, and thus also the ones that are easiest to predict correctly. Winner-take-all link assignment methods, such as the competitive linking algorithm, can prevent links based on indirect associations (see Section 4.2), thereby leveraging their accuracy on the more confident links to raise the accuracy of the less confident links. For example, suppose that u_1 and u_2 co-occur with v_1 and v_2 in the training data, and the model estimates $score(\mathbf{u_1}, \mathbf{v_1}) = .05$, $score(\mathbf{u_1}, \mathbf{v_2}) = .02$, and $score(\mathbf{u_2}, \mathbf{v_2}) = .01$. According to the one-to-one assumption, (u_1, v_2) is an indirect association and the correct translation of v_2 is u_2. To the extent that the one-to-one assumption is valid, it reduces the probability of spurious links for the rarer words. The more incorrect candidate translations can be eliminated for a given rare word, the more likely the correct translation is to be found. So, the probability of a correct match for a rare word is proportional to the fraction of words around it that can be linked with higher confidence. This fraction is largely determined by two bitext properties: the distribution of word frequencies, and the distribution of co-occurrence counts. Melamed (to appear) explores these properties in greater depth.

5.1.3 Step 3: Reestimation of the Model Parameters. Method A reestimates the *score* parameters as the logarithm of the *trans* parameters. The competitive linking algorithm only cares about the *relative* magnitudes of the various $score(\mathbf{u}, \mathbf{v})$. However, Equation 26 is a sum rather than a product, so I scale the *trans* parameters logarithmically, to be consistent with its probabilistic interpretation:

$$score_A(\mathbf{u}, \mathbf{v}) = \log trans(\mathbf{u}, \mathbf{v}) \tag{28}$$

5.2 Method B: Improved Estimation Using an Explicit Noise Model

Yarowsky (1993, 271) has shown that "for several definitions of sense and collocation, an ambiguous word has only one sense in a given collocation with a probability of 90–99%." In other words, a single contextual clue can be a highly reliable indicator of a word's sense. One of the definitions of "sense" studied by Yarowsky was a word token's translation in the other half of a bitext. For example, the English word *sentence* may be considered to have two senses, corresponding to its French translations *peine* (judicial sentence) and *phrase* (grammatical sentence). If a token of *sentence* occurs in the vicinity of a word like *jury* or *prison*, then it is far more likely to be translated as *peine* than as *phrase*. "In the vicinity of" is one kind of collocation. Co-occurrence

11 The competitive linking algorithm can be generalized to stop searching before the number of possible assignments is reduced to one, at which point the link counts can be computed as probabilistically weighted averages over the remaining assignments. I use this method to resolve ties.

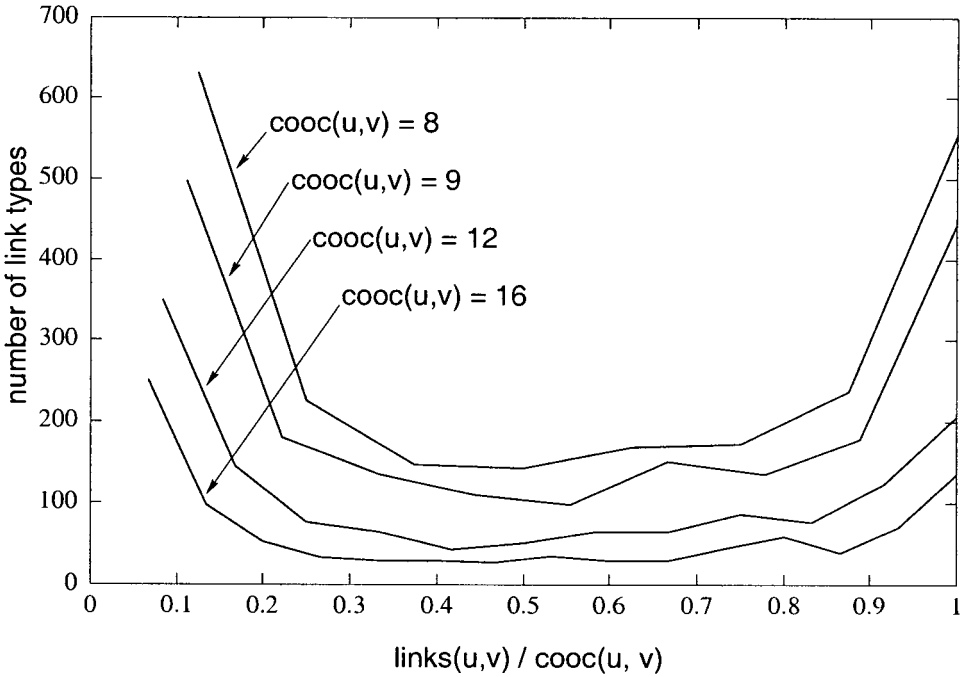


Figure 2
The ratio $links(\mathbf{u},\mathbf{v})/cooc(\mathbf{u},\mathbf{v})$, for several values of $cooc(\mathbf{u},\mathbf{v})$.

in bitext space is another kind of collocation. If each word's translation is treated as a sense tag (Resnik and Yarowsky 1997), then "translational" collocations have the unique property that the collocate and the word sense are one and the same!

Method B exploits this property under the hypothesis that "one sense per collocation" holds for translational collocations. This hypothesis implies that if **u** and **v** are *possible* mutual translations, and a token u co-occurs with a token v in the bitext, then with very high probability the pair (u, v) was generated from the same concept and should be linked. To test this hypothesis, I ran one iteration of Method A on 300,000 aligned sentence pairs from the Canadian Hansards bitext. I then plotted the ratio $\frac{links(\mathbf{u},\mathbf{v})}{cooc(\mathbf{u},\mathbf{v})}$ for several values of $cooc(\mathbf{u},\mathbf{v})$ in Figure 2. The curves show that the ratio $\frac{links(\mathbf{u},\mathbf{v})}{cooc(\mathbf{u},\mathbf{v})}$ tends to be either very high or very low. This bimodality is not an artifact of the competitive linking process, because in the first iteration, linking decisions are based only on the initial similarity metric.

Information about how often words co-occur without being linked can be used to bias the estimation of translation model parameters. The smaller the ratio $\frac{links(\mathbf{u},\mathbf{v})}{cooc(\mathbf{u},\mathbf{v})}$, the more likely it is that **u** and **v** are *not* mutual translations, and that links posited between tokens of **u** and **v** are noise. The bias can be implemented via auxiliary parameters that model the curve illustrated in Figure 2. The competitive linking algorithm creates all the links of a given type independently of each other.[12] So, the distribution of the number $links(\mathbf{u},\mathbf{v})$ of links connecting word types **u** and **v** can be modeled by a binomial distribution with parameters $cooc(\mathbf{u},\mathbf{v})$ and $p(\mathbf{u},\mathbf{v})$. $p(\mathbf{u},\mathbf{v})$ is the probability

12 Except for the case when multiple tokens of the same word type occur near each other, which I hereby sweep under the carpet.

Table 3
Variables used to describe Method B.

$links(\mathbf{u}, \mathbf{v})$	=	the number of times that **u** and **v** are hypothesized to co-occur as mutual translations
$B(k\|n, p)$	=	probability of k being generated from a binomial distribution with parameters n and p
λ^+	=	probability of a link given mutual translations
λ^-	=	probability of a link given not mutual translations
λ	=	probability of a link
τ	=	probability of mutual translations
K	=	total number of links in the bitext
N	=	total number of co-occurrences in the bitext

that **u** and **v** will be linked when they co-occur. There is never enough data to robustly estimate each p parameter separately. Instead, I shall model all the p's with just two parameters. For **u** and **v** that are mutual translations, $p(\mathbf{u}, \mathbf{v})$ will average to a relatively high probability, which I will call λ^+. for **u** and **v** that are not mutual translations, $p(\mathbf{u}, \mathbf{v})$ will average to a relatively low probability, which I will call λ^-. λ^+ and λ^- correspond to the two peaks of the distribution $\frac{links(\mathbf{u},\mathbf{v})}{cooc(\mathbf{u},\mathbf{v})}$, which is illustrated in Figure 2. The two parameters can also be interpreted as the rates of true and false positives. If the translation in the bitext is consistent and the translation model is accurate, then λ^+ will be close to one and λ^- will be close to zero.

To find the most likely values of the auxiliary parameters λ^+ and λ^-, I adopt the standard method of maximum likelihood estimation, and find the values that maximize the probability of the link frequency distributions, under the usual independence assumptions:

$$\Pr(links|model) = \prod_{\mathbf{u},\mathbf{v}} \Pr(links(\mathbf{u}, \mathbf{v})|cooc(\mathbf{u}, \mathbf{v}), \lambda^+, \lambda^-) \tag{29}$$

Table 3 summarizes the variables involved in this auxiliary estimation process.

The factors on the right-hand side of Equation 29 can be written explicitly with the help of a mixture coefficient. Let τ be the probability that an arbitrary co-occurring pair of word types are mutual translations. Let $B(k|n, p)$ denote the probability that k links are observed out of n co-occurrences, where k has a binomial distribution with parameters n and p. Then the probability that word types **u** and **v** will be linked $links(\mathbf{u}, \mathbf{v})$ times out of $cooc(\mathbf{u}, \mathbf{v})$ co-occurrences is a mixture of two binomials:

$$\begin{aligned} \Pr(links(\mathbf{u}, \mathbf{v})|cooc(\mathbf{u}, \mathbf{v}), \lambda^+, \lambda^-) &= \tau B(links(\mathbf{u}, \mathbf{v})|cooc(\mathbf{u}, \mathbf{v}), \lambda^+) \\ &\quad + (1 - \tau) B(links(\mathbf{u}, \mathbf{v})|cooc(\mathbf{u}, \mathbf{v}), \lambda^-). \end{aligned} \tag{30}$$

One more variable allows us to express τ in terms of λ^+ and λ^-: Let λ be the probability that an arbitrary co-occuring pair of word tokens will be linked, regardless of whether they are mutual translations. Since τ is constant over all word types, it also represents the probability that an arbitrary co-occurring pair of word *tokens* are mutual translations. Therefore,

$$\lambda = \tau\lambda^+ + (1 - \tau)\lambda^-. \tag{31}$$

λ can also be estimated empirically. Let K be the total number of links in the bitext

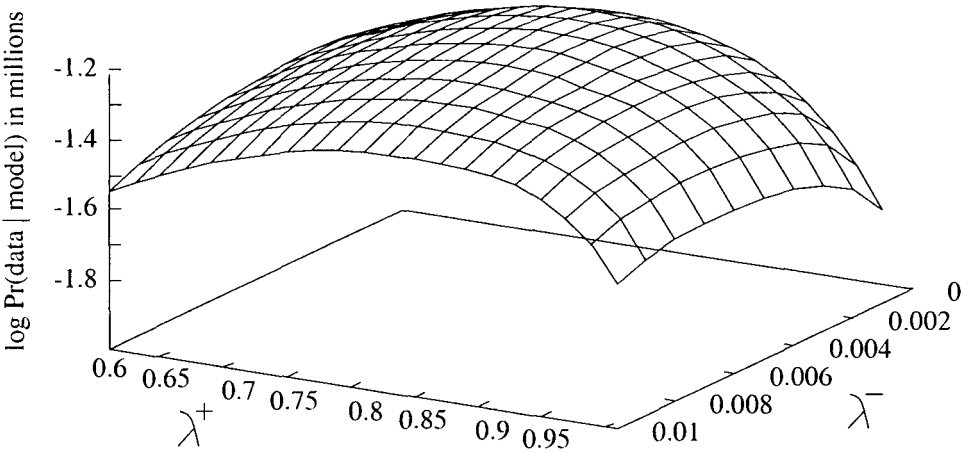


Figure 3
Pr(*links*|*model*), as given in Equation 29, has only one global maximum in the region of interest, where $1 > \lambda^+ > \lambda > \lambda^- > 0$.

and let N be the total number of word token pair co-occurrences:

$$K = \sum_{\mathbf{u},\mathbf{v}} links(\mathbf{u}, \mathbf{v}), \tag{32}$$

$$N = \sum_{\mathbf{u},\mathbf{v}} cooc(\mathbf{u}, \mathbf{v}). \tag{33}$$

By definition,

$$\lambda = K/N. \tag{34}$$

Equating the right-hand sides of Equations 31 and 34 and rearranging the terms, we get:

$$\tau = \frac{K/N - \lambda^-}{\lambda^+ - \lambda^-}. \tag{35}$$

Since τ is now a function of λ^+ and λ^-, only the latter two variables represent degrees of freedom in the model.

In the preceding equations, either **u** or **v** can be NULL. However, the number of times that a word co-occurs with NULL is not an observable feature of bitexts. To make sense of co-occurrences with NULL, we can view co-occurrences as *potential* links and $cooc(\mathbf{u}, \mathbf{v})$ as the maximum number of times that tokens of **u** and **v** might be linked. From this point of view, $cooc(\mathbf{u}, \text{NULL})$ should be set to the unigram frequency of **u**, since each token of **u** represents one potential link to NULL. Similarly for $cooc(\text{NULL}, \mathbf{v})$. These co-occurrence counts should be summed together with all the others in Equation 33.

The probability function expressed by Equations 29 and 30 may have many local maxima. In practice, these local maxima are like pebbles on a mountain, invisible at low resolution. I computed Equation 29 over various combinations of λ^+ and λ^- after one iteration of Method A over 300,000 aligned sentence pairs from the Canadian Hansard bitext. Figure 3 illustrates that the region of interest in the parameter space, where $1 > \lambda^+ > \lambda > \lambda^- > 0$, has only one dominant global maximum. This global maximum can be found by standard hill-climbing methods, as long as the step size is large enough to avoid getting stuck on the pebbles.

Given estimates for λ^+ and λ^-, we can compute $B(links(\mathbf{u},\mathbf{v})|cooc(\mathbf{u},\mathbf{v}),\lambda^+)$ and $B(links(\mathbf{u},\mathbf{v})|cooc(\mathbf{u},\mathbf{v}),\lambda^-)$ for each occurring combination of *links* and *cooc* values. These are the probabilities that $links(\mathbf{u},\mathbf{v})$ links were generated out of $cooc(\mathbf{u},\mathbf{v})$ possible links by a process that generates correct links and by a process that generates incorrect links, respectively. The ratio of these probabilities is the likelihood ratio in favor of the types $\mathbf{u}$ and $\mathbf{v}$ being possible mutual translations, for all $\mathbf{u}$ and $\mathbf{v}$:

$$score_B(\mathbf{u},\mathbf{v}) = \log \frac{B(links(\mathbf{u},\mathbf{v})|cooc(\mathbf{u},\mathbf{v}),\lambda^+)}{B(links(\mathbf{u},\mathbf{v})|cooc(\mathbf{u},\mathbf{v}),\lambda^-)}. \tag{36}$$

Method B differs from Method A only in its redefinition of the *score* function in Equation 36. The auxiliary parameters λ^+ and λ^- and the noise model that they represent can be employed the same way in translation models that are not based on the one-to-one assumption.

5.3 Method C: Improved Estimation Using Preexisting Word Classes

In Method B, the estimation of the auxiliary parameters λ^+ and λ^- depends only on the overall distribution of co-occurrence counts and link frequencies. All word pairs that co-occur the same number of times and are linked the same number of times are assigned the same *score*. More accurate models can be induced by taking into account various features of the linked tokens. For example, frequent words are translated less consistently than rare words (Catizone, Russell, and Warwick 1989). To account for these differences, we can estimate separate values of λ^+ and λ^- for different ranges of $cooc(\mathbf{u},\mathbf{v})$. Similarly, the auxiliary parameters can be conditioned on the linked parts of speech. A kind of word order correlation bias can be effected by conditioning the auxiliary parameters on the relative positions of linked word tokens in their respective texts. Just as easily, we can model link types that coincide with entries in an on-line bilingual dictionary separately from those that do not (cf. Brown et al. 1993). When the auxiliary parameters are conditioned on different link classes, their optimization is carried out separately for each class:

$$score_C(\mathbf{u},\mathbf{v}|Z = class(\mathbf{u},\mathbf{v})) = \log \frac{B(links(\mathbf{u},\mathbf{v})|cooc(\mathbf{u},\mathbf{v}),\lambda_Z^+)}{B(links(\mathbf{u},\mathbf{v})|cooc(\mathbf{u},\mathbf{v}),\lambda_Z^-)}. \tag{37}$$

Section 6.1.1 describes the link classes used in the experiments below.

6. Evaluation

6.1 Evaluation at the Token Level

This section compares translation model estimation methods A, B, and C to each other and to Brown et al.'s (1993b) Model 1. To reiterate, Model 1 is based on co-occurrence information only; Method A is based on the one-to-one assumption; Method B adds the "one sense per collocation" hypothesis to Method A; Method C conditions the auxiliary parameters of Method B on various word classes. Whereas Methods A and B and Model 1 were fully specified in Section 4.3.1 and Section 5, the latter section described a variety of features on which Method C might classify links. For the purposes of the experiments described in this article, Method C employed the simple classification in Table 4 for both languages in the bitext. All classification was performed by table lookup; no context-aware part-of-speech tagger was used. In particular, words that were ambiguous between open classes and closed classes were always deemed to be in the closed class. The only language-specific knowledge involved in this classification

Table 4
Word classes used by Method C for the experiments described in this article. Link classes were constructed by taking the cross-product of the word classes.

Class Code	Description
EOS	End-Of-Sentence punctuation
EOP	End-Of-Phrase punctuation, such as commas and colons
SCM	Subordinate Clause Markers, such as " and (
SYM	Symbols, such as ˜ and *
NU	the NULL word, in a class by itself
C	Content words: nouns, adjectives, adverbs, non-auxiliary verbs
F	all other words, i.e., function words

method is the list of function words in class F. Certainly, more sophisticated word classification methods could produce better models, but even the simple classification in Table 4 should suffice to demonstrate the method's potential.

6.1.1 Experiment 1. Until now, translation models have been evaluated either subjectively (e.g. White and O'Connell 1993) or using relative metrics, such as perplexity with respect to other models (Brown et al. 1993b). Objective and more accurate tests can be carried out using a "gold standard." I hired bilingual annotators to link roughly 16,000 corresponding words between on-line versions of the Bible in French and English. This bitext was selected to facilitate widespread use and standardization (see Melamed [1998c] for details). The entire Bible bitext comprised 29,614 verse pairs, of which 250 verse pairs were hand-linked using a specially developed annotation tool. The annotation style guide (Melamed 1998b) was based on the intuitions of the annotators, so it was not biased towards any particular translation model. The annotation was replicated five times by seven different annotators.

Each of the four methods was used to estimate a word-to-word translation model from the 29,614 verse pairs in the Bible bitext. All methods were deemed to have converged when less than .0001 of the translational probability distribution changed from one iteration to the next. The links assigned by each of methods A, B, and C in the last iteration were normalized into joint probability distributions using Equation 19. I shall refer to these joint distributions as Model A, Model B, and Model C, respectively. Each of the joint probability distributions was further normalized into two conditional probability distributions, one in each direction. Since Model 1 is inherently directional, its conditional probability distributions were estimated separately in each direction, instead of being derived from a joint distribution.

The four models' predictions were compared to the gold standard annotations. Each model guessed one translation (either stochastically or deterministically, depending on the task) for each word on one side of the gold standard bitext. Therefore, precision = recall here, and I shall refer to the results simply as "percent correct." The accuracy of each model was averaged over the two directions of translation: English to French and French to English. The five-fold replication of annotations in the test data enabled computation of the statistical significance of the differences in model accuracy. The statistical significance of all results in this section was measured at the $\alpha = .05$ level, using the Wilcoxon signed ranks test. Although the models were evaluated on part of the same bitext on which they were trained, the evaluations were with respect to the translational equivalence relation hidden in this bitext, not with respect to any of the bitext's visible features. Such testing on training data is standard practice for

unsupervised learning algorithms, where the objective is to compare several methods. Of course, performance would degrade on previously unseen data.

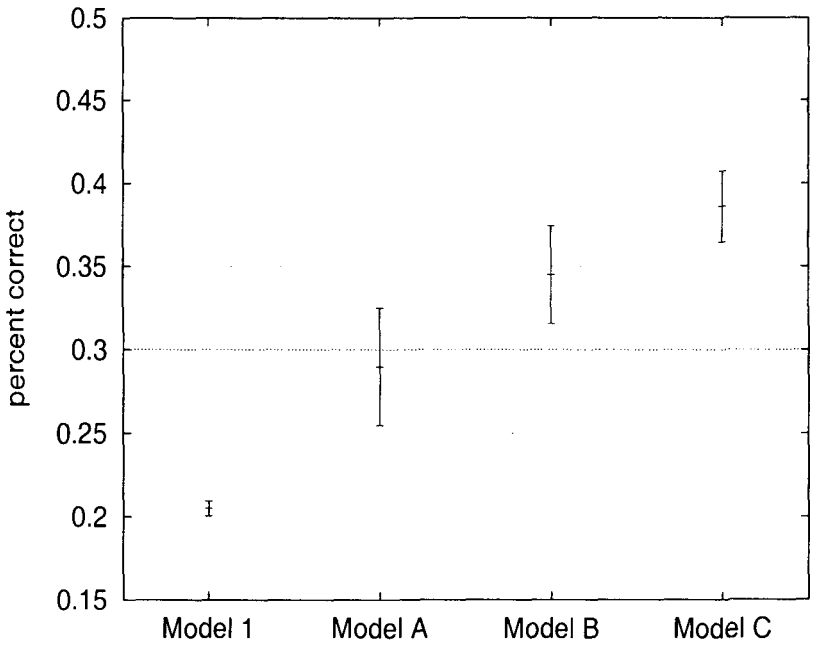
In addition to the different translation models, there were two other independent variables in the experiment: method of translation and whether function words were included. Some applications, such as query translation for CLIR, don't care about function words. To get a sense of the relative effectiveness of the different translation model estimation methods when function words are taken out of the equation, I removed from the gold standard all link tokens where one or both of the linked words were closed-class words. Then, I removed all closed-class words (including nonalphabetic symbols) from the models and renormalized the conditional probabilities.

The method of translation was either **single-best** or **whole distribution**. Single-best translation is the kind that somebody might use to get the gist of a foreign-language document. The input to the task was one side of the gold standard bitext. The output was the model's single best guess about the translation of each word in the input, together with the input word. In other words, each model produced link tokens consisting of input words and their translations. For some applications, it is insufficient to guess only the single most likely translation of each word in the input. The model is expected to output the whole distribution of possible translations for each input word. This distribution is then combined with other distributions that are relevant to the application. For example, for cross-language information retrieval, the translational distribution can be combined with the distribution of term frequencies. For statistical machine translation, the translational distribution can be decoded with a source language model (Brown et al. 1988; Al-Onaizan et al. 1999). To predict how the different models might perform in such applications, the whole distribution task was to generate a whole set of links from each input word, weighted according to the probability assigned by the model to each of the input word's translations. Each model was tested on this task with and without function words.

The mean results are plotted in Figures 4 and 5 with 95% confidence intervals. All four graphs in these figures are on the same scale to facilitate comparison. On both tasks involving the entire vocabulary, each of the biases presented in this article improves the efficiency of modeling the available training data. When closed-class words were ignored, Model 1 performed better than Method A, because open-class words are more likely to violate the one-to-one assumption. However, the explicit noise model in Methods B and C boosted their scores significantly higher than Model 1 and Method A. Method B was better than Method C at choosing the single best open-class links, and the situation was reversed for the whole distribution of open-class links. However, the differences in performance between these two methods were tiny on the open-class tasks, because they left only two classes for Method C to distinguish: content words and NULLs. Most of the scores on the whole distribution task were lower than their counterparts on the single-best translation task, because it is more difficult for any statistical method to correctly model the less common translations. The "best" translations are usually the most common.

6.1.2 Experiment 2. To study how the benefits of the various biases vary with training corpus size, I evaluated Models A, B, C, and 1 on the whole distribution translation task, after training them on three different-size subsets of the Bible bitext. The first subset consisted of only the 250 verse pairs in the gold standard. The second subset included these 250 plus another random sample of 2,250 for a total of 2,500, an order of magnitude larger than the first subset. The third subset contained all 29,614 verse pairs in the Bible bitext, roughly an order of magnitude larger than the second subset. All models were compared to the five gold standard annotations, and the scores were

(a)



(b)

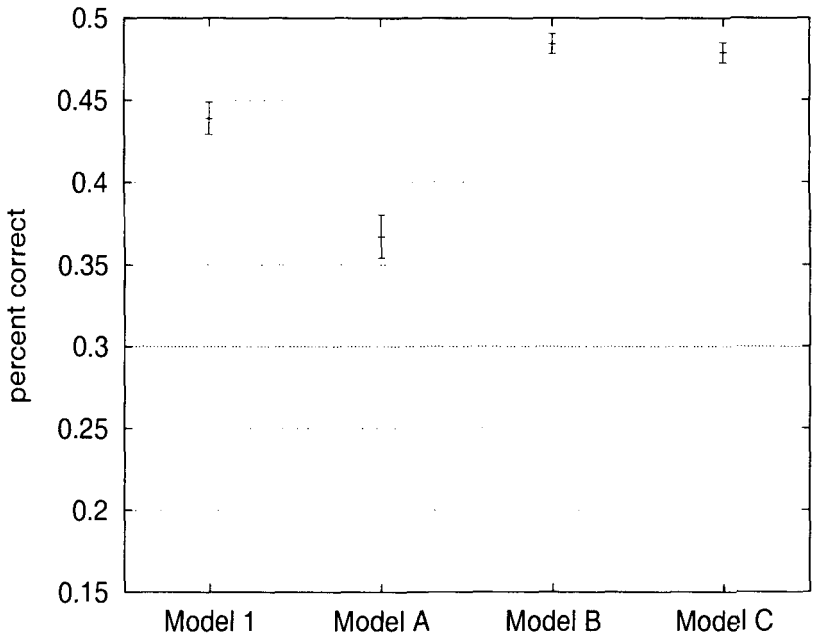


Figure 4
Comparison of model performance on single-best translation task. (a) All links; (b) open-class links only.

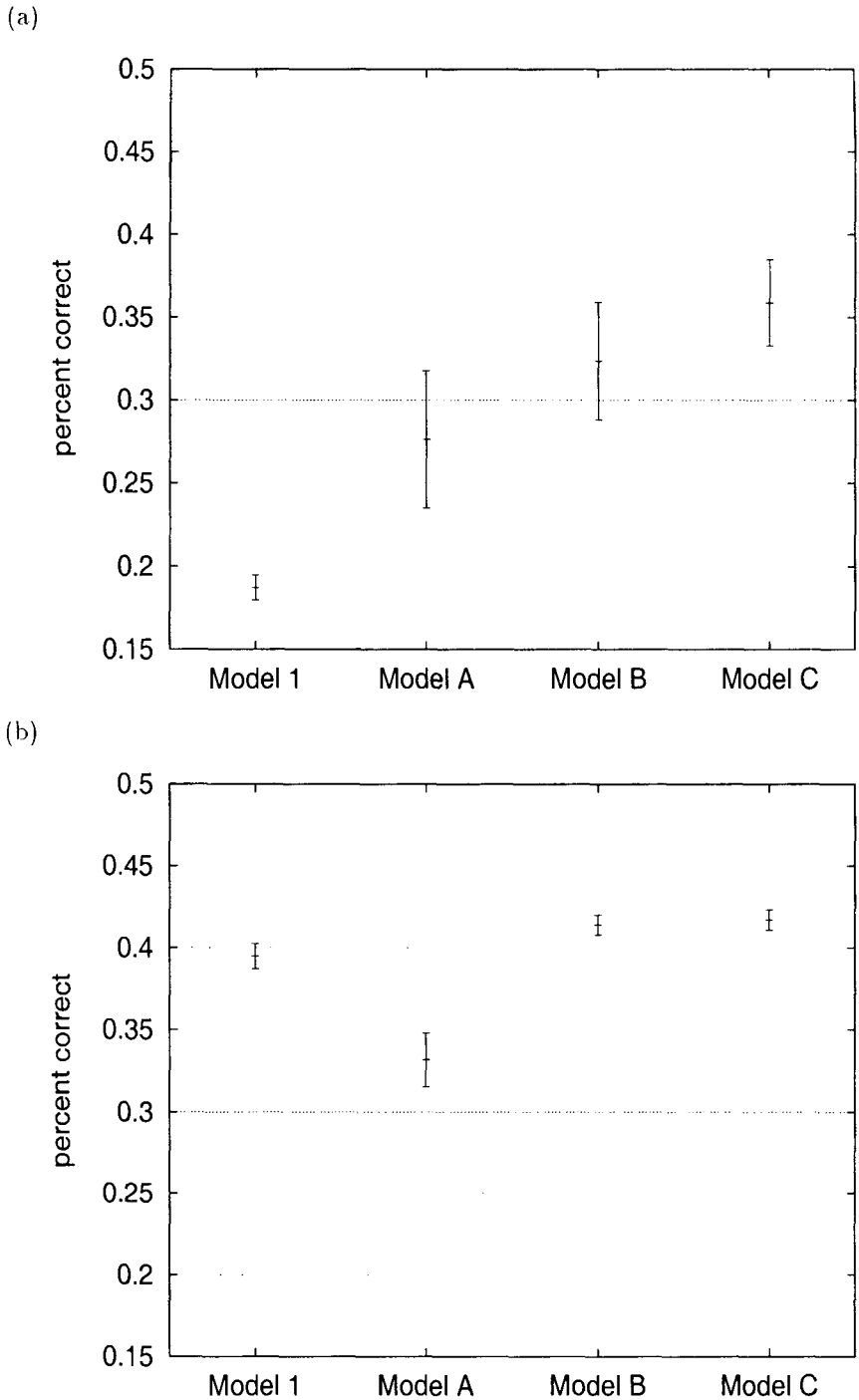


Figure 5
Comparison of model performance on whole distribution task. (a) All links; (b) open-class links only.

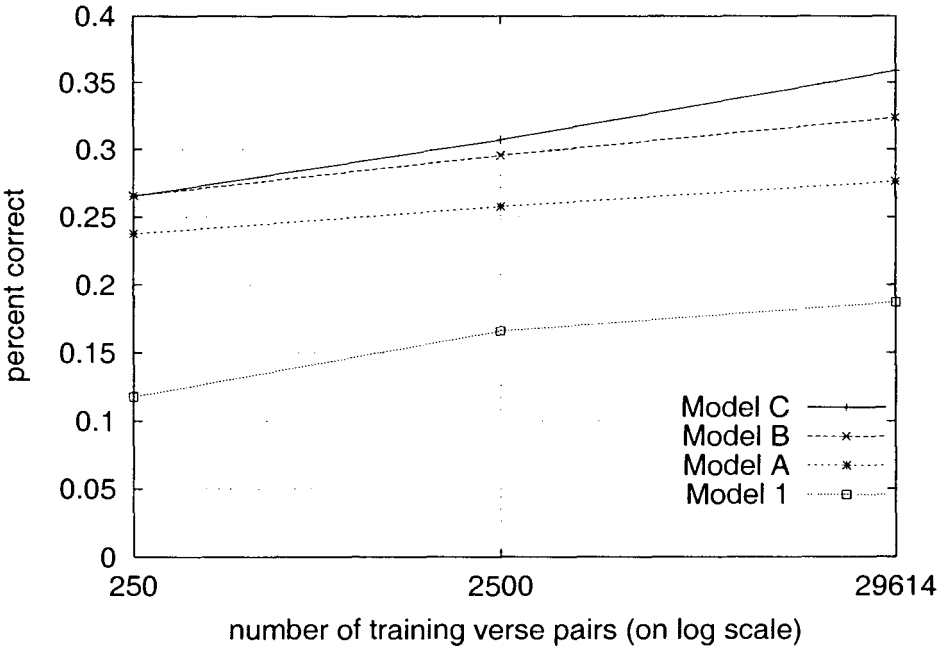


Figure 6
Effects of training set size on model accuracy on the whole distribution task.

averaged over the two directions of translation, as before. Again, because the total probability assigned to all translations for each source word was one, precision = recall = percent correct on this task. The mean scores over the five gold standard annotations are graphed in Figure 6, where the right edge of the figure corresponds to the means of Figure 5(a). The figure supports the hypothesis in Melamed (to appear, Chapter 7) that the biases presented in this article are even more valuable when the training data are more sparse. The one-to-one assumption is useful, even though it forces us to use a greedy approximation to maximum likelihood. In relative terms, the advantage of the one-to-one assumption is much more pronounced on smaller training sets. For example, Model A is 102% more accurate than Model 1 when trained on only 250 verse pairs. The explicit noise model buys a considerable gain in accuracy across all sizes of training data, as do the link classes of Model C. In concert, when trained and tested only on the gold standard test set, the three biases outperformed Model 1 by up to 125%. This difference is even more significant given the absolute performance ceiling of 82% established by the interannotator agreement rates on the gold standard.

6.2 Evaluation at the Type Level

An important application of statistical translation models is to help lexicographers compile bilingual dictionaries. Dictionaries are written to answer the question, "What are the possible translations of X?" This is a question about link types, rather than about link tokens.

Evaluation by link type is a thorny issue. Human judges often disagree about the degree to which context should play a role in judgments of translational equivalence. For example, the *Harper-Collins French Dictionary* (Cousin et al. 1990) gives the following French translations for English *appoint: nommer, engager, fixer, désigner*. Likewise, most

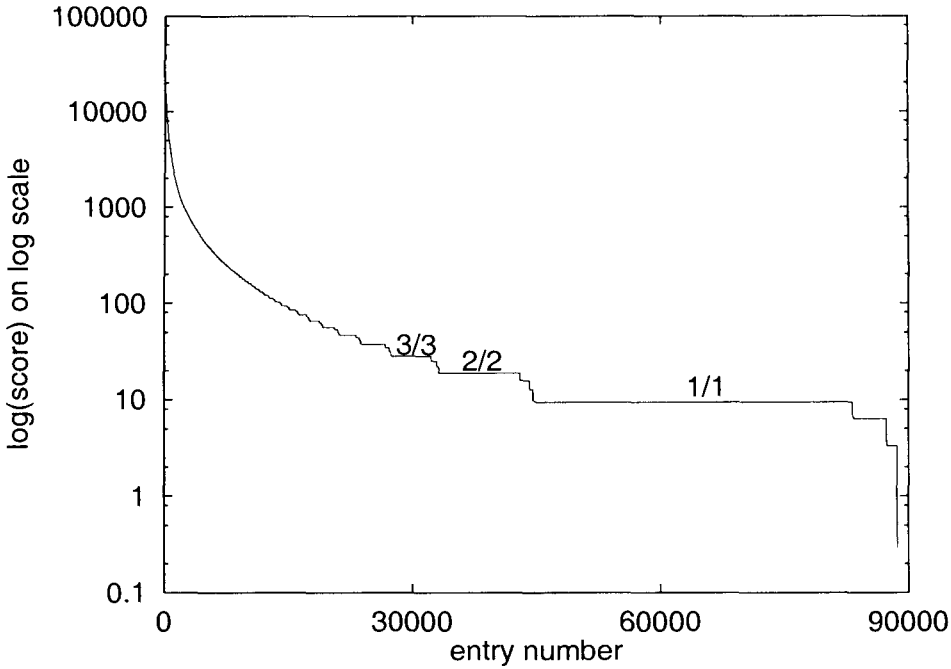


Figure 7
Distribution of link type scores. The long plateaus correspond to the most common combinations of $\frac{links(u,v)}{cooc(u,v)}$: 1/1, 2/2, and 3/3.

lay judges would not consider *instituer* a correct French translation of *appoint*. In actual translations, however, when the object of the verb is *commission, task force, panel, etc.*, English *appoint* is usually translated into French as *instituer*. To account for this kind of context-dependent translational equivalence, link types must be evaluated with respect to the bitext whence they were induced.

I performed a post hoc evaluation of the link types produced by an earlier version of Method B (Melamed 1996b). The bitext used for this evaluation was the same aligned Hansards bitext used by Gale and Church (1991), except that I used only 300,000 aligned segment pairs to save time. The bitext was automatically pretokenized to delimit punctuation, English possessive pronouns, and French elisions. Morphological variants in both halves of the bitext were stemmed to a canonical form.

The link types assigned by the converged model were sorted by the scores in Equation 36. Figure 7 shows the distribution of these scores on a log scale. The log scale helps to illustrate the plateaus in the curve. The longest plateau represents the set of word pairs that were linked once out of one co-occurrence (1/1) in the bitext. All these word pairs were equally likely to be correct. The second-longest plateau resulted from word pairs that were linked twice out of two co-occurrences (2/2) and the third longest plateau is from word pairs that were linked three times out of three co-occurrences (3/3). As usual, the entries with higher scores were more likely to be correct. By discarding entries with lower scores, coverage could be traded for accuracy. This trade-off was measured at three points, representing cutoffs at the end of each of the three longest plateaus.

The traditional method of measuring coverage requires knowledge of the correct link types, which is impossible to determine without a gold standard. An approximate coverage measure can be based on the number of different words in the corpus. For

Table 5
Lexicon coverage at three different minimum score thresholds. The bitext contained 41,028 different English words and 36,314 different French words, for a total of 77,342.

Cutoff Plateau	Minimum Score	Total Lexicon Entries	English Words Represented	%	French Words Represented	%
3/3	28	32,274	14,299	35	13,409	37
2/2	18	43,075	18,533	45	17,133	47
1/1	9	88,633	36,371	89	33,017	91

lexicons extracted from corpora, perfect coverage implies at least one entry containing each word in the corpus. One-sided variants, which consider only source words, have also been used (Gale and Church 1991). Table 5 shows both the marginal (one-sided) and the combined coverage at each of the three cutoff points. It also shows the absolute number of (non-NULL) entries in each of the three lexicons. Of course, the size of automatically induced lexicons depends on the size of the training bitext. Table 5 shows that, given a sufficiently large bitext, the method can automatically construct translation lexicons with as many entries as published bilingual dictionaries.

The next task was to measure accuracy. It would have taken too long to evaluate every lexicon entry manually. Instead, I took five random samples (with replacement) of 100 entries each from each of the three lexicons. Each of the samples was first compared to a translation lexicon extracted from a machine-readable bilingual dictionary (Cousin et al. 1991). All the entries in the sample that appeared in the dictionary were assumed to be correct. I checked the remaining entries in all the samples by hand. To account for context-dependent translational equivalence, I evaluated the accuracy of the translation lexicons in the context of the bitext whence they were extracted, using a simple bilingual concordancer. A lexicon entry **(u,v)** was considered correct if *u* and *v* ever appeared as direct translations of each other in an aligned segment pair. That is, a link type was considered correct if any of its tokens were correct.

Direct translations come in different flavors. Most entries that I checked by hand were of the plain vanilla variety that you might find in a bilingual dictionary (entry type V). However, a significant number of words translated into a different part of speech (entry type P). For instance, in the entry (protection, protégé), the English word is a noun but the French word is an adjective. This entry appeared because *to have protection* is often translated as *être protégé* ('to be protected') in the bitext. The entry will never occur in a bilingual dictionary, but users of translation lexicons, be they human or machine, will want to know that translations often happen this way.

The evaluation of translation models at the word type level is complicated by the possibility of phrasal translations, such as *immédiatement* ↔ *right away*. All the methods being evaluated here produce models of translational equivalence between individual words only. How can we decide whether a single-word translation "matches" a phrasal translation? The answer lies in the observation that corpus-based lexicography usually involves a lexicographer. Bilingual lexicographers can work with bilingual concordancing software that can point them to instances of any link type induced from a bitext and display these instances sorted by their contexts (e.g. Simard, Foster, and Perrault 1993). Given an incomplete link type, the lexicographer can usually reconstruct the complete link type from the contexts in the concordance. For example, if the model proposes an equivalence between *immédiatement* and *right*, a bilingual concordance

Table 6
Distribution of different types of correct lexicon entries at varying levels of coverage (mean ± standard deviation).

Cutoff	Coverage	% Type V	% Type P	% Type I	Total % Accuracy
3/3	36%	89 ± 2.2	3.4 ± 0.5	7.6 ± 3.2	99.2 ± 0.8
2/2	46%	81 ± 3.0	8.0 ± 2.1	9.8 ± 1.8	99.0 ± 1.4
1/1	90%	82 ± 2.5	4.4 ± 0.5	6.0 ± 1.9	92.8 ± 1.1

can show the lexicographer that the model was really trying to capture the equivalence between *immédiatement* and *right away* or between *immédiatement* and *right now*. I counted incomplete entries in a third category (entry type I). Whether links in this category should be considered correct depends on the application.

Table 6 shows the distribution of correct lexicon entries among the types V, P and I. Figure 8 graphs the accuracy of the method against coverage, with 95% confidence intervals. The upper curve represents accuracy when incomplete links are considered correct, and the lower when they are considered incorrect. On the former metric, the method can generate translation lexicons with accuracy and coverage both exceeding 90%, as well as dictionary-size translation lexicons that are over 99% correct.

7. Conclusion

There are many ways to model translational equivalence and many ways to estimate translation models. "The mathematics of statistical machine translation" proposed by Brown et al. (1993b) are just one kind of mathematics for one kind of statistical trans-

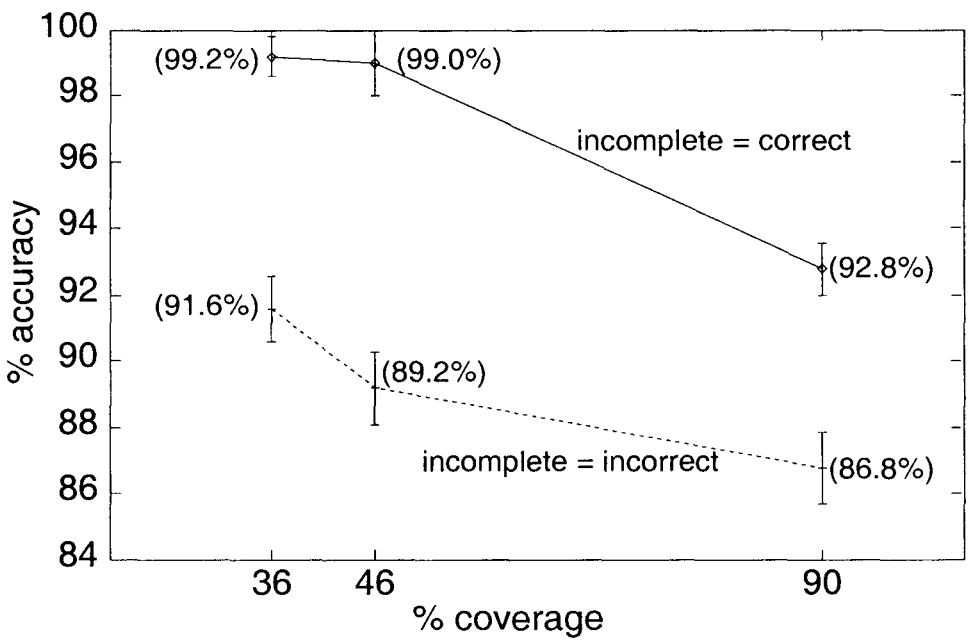


Figure 8
Translation lexicon accuracy with 95% confidence intervals at varying levels of coverage.

lation. In this article, I have proposed and evaluated new kinds of translation model biases, alternative parameter estimation strategies, and techniques for exploiting preexisting knowledge that may be available about particular languages and language pairs. On a variety of evaluation metrics, each infusion of knowledge about the problem domain resulted in better translation models.

Each innovation presented here opens the way for more research. Model biases can be mixed and matched with each other, with previously published biases like the word order correlation bias, and with other biases yet to be invented. The competitive linking algorithm can be generalized in various ways. New kinds of preexisting knowledge can be exploited to improve accuracy for particular language pairs or even just for particular bitexts. It is difficult to say where the greatest advances will come from. Yet, one thing is clear from our current vantage point: Research on empirical methods for modeling translational equivalence has not run out of steam, as some have claimed, but has only just begun.

Acknowledgments

Much of this research was performed at the Department of Computer and Information Science at the University of Pennsylvania, where it was supported by an equipment grant from Sun MicroSystems Laboratories and by ARPA Contract #N66001-94C-6043. Many thanks to my former colleagues at UPenn and to the anonymous reviewers for their insightful suggestions for improvement.

References

Abeillé, Anne, Yves Schabes, and Aravind K. Joshi. 1990. Using lexicalized tree adjoining grammars for machine translation. In *Proceedings of the 13th International Conference on Computational Linguistics*. Helsinki, Finland.

Ahuja, Ravindra K., Thomas L. Magnati, and James B. Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Al-Onaizan, Yaser, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Franz J. Och, David Purdy, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation. CLSP Technical Report. Baltimore, MD. Available at www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps

Brousseau, Julie, Caroline Drouin, George Foster, Pierre Isabelle, Roland Kuhn, Yves Normandin, and Pierre Plamondon. 1995. French speech recognition in an automatic dictation system for translators: The TransTalk project. In *Proceedings of EuroSpeech'95*, pages 193–196, Madrid, Spain.

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

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer and Surya Mohanty. 1993a. But dictionaries are data too. In *Proceedings of the ARPA HLT Workshop*, pages 202–205, Princeton, NJ.

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

Buckley, Chris. 1993. The importance of proper weighting methods. In *Proceedings of the DARPA Workshop on Human Language Technology*, pages 349–352, Princeton, NJ.

Candito, Marie-Hélène. 1998. Building parallel LTAG for French and Italian. In *COLING-ACL '98: 36 Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 211–217, Montreal, Canada.

Catizone, Roberta, Graham Russell, and Susan Warwick. 1989. Deriving translation data from bilingual texts. In *Proceedings of the First International Lexical Acquisition Workshop*. Detroit, MI.

Church, Kenneth W., and Eduard H. Hovy. 1993. Good applications for crummy machine translation. *Machine Translation 8*.

Cousin, Pierre-Henri, Lorna Sinclair, Jean-François Allain, and Catherine E. Love. 1990. *The Harper Collins French Dictionary*. Harper Collins Publishers,

New York, NY.
Cousin, Pierre-Henri, Lorna Sinclair, Jean-François Allain, and Catherine E. Love. 1991. *The Collins Paperback French Dictionary*. Harper Collins Publishers, Glasgow.
Dagan, Ido, Kenneth W. Church, and William A. Gale. 1993. Robust word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8, Columbus, OH.
Daille, Béatrice, Éric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. *Proceedings of the 15th International Conference on Computational Linguistics*, pages 515–521, Kyoto, Japan.
Deming, W. Edwards, and Frederick F. Stephan. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11:427–444.
Dempster, Arthur P., N. M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38.
Dorr, Bonnie J. 1992. The use of lexical semantics in interlingual machine translation. *Machine Translation*, 7(3):135–193.
Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74.
Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
Foster, George, Pierre Isabelle, and Pierre Plamondon. 1996. Word completion: A first step toward target-text mediated IMT. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 394–399, Copenhagen, Denmark.
Fung, Pascale. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Meeting*, pages 236–243, Boston, MA. Association for Computational Linguistics.
Gale, William A., and Kenneth W. Church. 1991. Identifying word correspondences in parallel texts. *Proceedings of the DARPA SNL Workshop*, pages 152–157, Asilomar, CA.
Gale, William A., and Geoff Sampson. 1995. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2:217–237. Swets & Zeitlinger Publishers, Sassenheim, The Netherlands.
Hiemstra, Djoerd. 1996. Using Statistical Methods to Create a Bilingual Dictionary. Masters thesis, University of Twente, The Netherlands.
Hiemstra, Djoerd. 1998. Multilingual domain modeling in twenty-one: Automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Proceedings of the Eighth meeting of Computational Linguistics in the Netherlands (CLIN)*, pages 41–58.
Kumano, Akira, and Hideki Hirakawa. 1994. Building an MT dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 76–81, Kyoto, Japan.
McCarley, J. Scott. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting*, pages 208–214, College Park, MD. Association for Computational Linguistics.
Macklovitch, Elliott. 1994. Using bi-textual alignment for translation validation: The TransCheck system. In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pages 157–168. Columbia, MD.
Melamed, I. Dan. 1995. Automatic evaluation and uniform filter cascades for inducing *N*-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 184–198, Cambridge, MA.
Melamed, I. Dan. 1996a. Automatic detection of omissions in translations. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 764–769, Copenhagen, Denmark.
Melamed, I. Dan. 1996b. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas*, pages 125–134, Montreal, Canada.
Melamed, I. Dan. 1998a. Models of co-occurrence. Institute for Research in Cognitive Science Technical Report #98-05. University of Pennsylvania, Philadelphia, PA.
Melamed, I. Dan. 1998b. Annotation style guide for the blinker project. Institute for Research in Cognitive Science Technical Report #98-06. University of Pennsylvania, Philadelphia, PA.
Melamed, I. Dan. 1998c. Manual annotation of translational equivalence: The blinker project. Institute for Research in Cognitive

Science Technical Report #98-07. University of Pennsylvania, Philadelphia, PA.

Melamed, I. Dan. To appear. *Empirical Methods for Exploiting Parallel Texts*, MIT Press.

Nerbonne, John, Lauri Karttunen, Elena Paskaleva, Gabor Proszeky, and Tiit Roosmaa. 1997. Reading more into foreign languages. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 135–138, Washington, DC.

Oard, Douglas W. 1997. Adaptive filtering of multilingual document streams. In *Proceedings of the 5th RIAO Conference on Computer-Assisted Information Retrieval*, pages 233–253, Montreal, Canada.

Resnik, Philip. 1997. Evaluating multilingual gisting of Web pages. In *Proceedings of the AAAI Symposium on Natural Language Processing for the World Wide Web*. Stanford, CA.

Resnik, Philip, and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics*, pages 79–86, Washington, DC.

Shieber, Stuart. 1994. Restricting the weak-generative capacity of synchronous tree-adjoining grammars. *Computational Intelligence*, 10(4):371–385.

Simard, Michel, George F. Foster, and François Perrault. 1993. TransSearch: A bilingual concordance tool. Centre d'innovation en technologies de l'information, Laval, Canada.

Svartvik, Jan. 1992. *Directions in Corpus Linguistics*. Mouton de Gruyter, Berlin.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

Piek, Vossen, editor. 1998. *Eurowordnet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

White, John S., and Theresa A. O'Connell. 1993. Evaluation of machine translation. In *Proceedings of the ARPA HLT Workshop*, pages 206–210, Princeton, NJ.

Wu, Dekai, and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213, Columbia, MD.

Yarowsky, David. 1993. One sense per collocation. In *Proceedings of the DARPA Workshop on Human Language Technology*, pages 266–271, Princeton, NJ.