But Dictionaries Are Data Too

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Meredith J. Goldsmith, Jan Hajic, Robert L. Mercer, and Surya Mohanty

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598

ABSTRACT

Although empiricist approaches to machine translation depend vitally on data in the form of large bilingual corpora, bilingual dictionaries are also a source of information. We show how to model at least a part of the information contained in a bilingual dictionary so that we can treat a bilingual dictionary and a bilingual corpus as two facets of a unified collection of data from which to extract values for the parameters of a probabilistic machine translation system. We give an algorithm for obtaining maximum likelihood estimates of the parameters of a probabilistic model from this combined data and we show how these parameters are affected by inclusion of the dictionary for some sample words.

There is a sharp dichotomy today between rationalist and empiricist approaches to machine translation: rationalist systems are based on information cajoled fact by reluctant fact from the minds of human experts; empiricist systems are based on information gathered wholesale from data. The data most readily digested by our translation system is from bilingual corpora, but bilingual dictionaries are data too, and in this paper we show how to weave information from them into the fabric of our statistical model of the translation process.

When a lexicographer creates an entry in a bilingual dictionary, he describes in one language the meaning and use of a word from another language. Often, he includes a list of simple translations. For example, the entry for *disingenuousness* in the Harper-Collins Robert French Dictionary [1] lists the translations *déloyauté*, *manque de sincérité*, and *fourberie*. In constructing such a list, the lexicographer gathers, either through introspection or extrospection, instances in which *disingenuousness* has been used in various ways and records those of the different translations that he deems of sufficient importance. Although a dictionary is more than just a collection of lists, we will concentrate here on that portion of it that is made up of lists.

We formalize an intuitive account of lexicographic behavior as follows. We imagine that a lexicographer, when constructing an entry for the English word or phrase $\mathbf{e}$, first chooses a random size s, and then selects at random a sample of s instances of the use of $\mathbf{e}$, each with its French translation. We imagine, further, that he includes in his entry for $\mathbf{e}$ a list consisting of all of the translations that occur at least once in his random sample. The probability that he will, in this way, obtain the list $\mathbf{f}_1, \ldots, \mathbf{f}_m$, is

$$\Pr(\mathbf{f}_1, \ldots, \mathbf{f}_m|e) = \tag{1}$$

$$\sum_s \sum_{s_1>0} \cdots \sum_{s_m>0} \binom{s}{s_1 \cdots s_m} \Pr(s|\mathbf{e}) \prod_{i=1}^m \Pr(\mathbf{f}_i|\mathbf{e})^{s_i},$$

where $\Pr(\mathbf{f}_i|\mathbf{e})$ is the probability from our statistical model that the phrase $\mathbf{f}_i$ occurs as a translation of $\mathbf{e}$, and $\Pr(s|\mathbf{e})$ is the probability that the lexicographer chooses to sample s instances of $\mathbf{e}$. The multinomial coefficient is defined by

$$\binom{s}{s_1 \cdots s_k} = \frac{s!}{s_1! \cdots s_k!}, \tag{2}$$

and satisfies the recursion

$$\binom{s}{s_1 \cdots s_k} = \binom{s}{s_k} \binom{s - s_k}{s_1 \cdots s_{k-1}} \tag{3}$$

where $\binom{s}{s_k}$ is the usual binomial coefficient.

In general, the sum in Equation (1) cannot be evaluated in closed form, but we can organize an efficient calculation of it as follows. Let

$$\alpha(\sigma,\mu) = \sum_{s_1>0} \cdots \sum_{s_\mu>0} \binom{\sigma}{s_1 \cdots s_\mu} \prod_{i=1}^{\mu} p(\mathbf{f}_i|\mathbf{e})^{s_i}. \quad (4)$$

Clearly,

$$p(\mathbf{f}_1,\cdots,\mathbf{f}_m|\mathbf{e}) = \sum_s p(s|\mathbf{e})\alpha(s,m). \quad (5)$$

Using Equation (3), it is easy to show that

$$\alpha(\sigma,\mu) = \sum_{\kappa=1}^{\sigma-\mu+1} \binom{\sigma}{\kappa} p(\mathbf{f}_\mu|\mathbf{e})^\kappa \alpha(\sigma-\kappa,\mu-1), \quad (6)$$

and therefore, we can compute $p(\mathbf{f}_1,\cdots,\mathbf{f}_m|\mathbf{e})$ in time proportional to s^2m. By judicious use of thresholds, even this can be substantially reduced.

In the special case that $\Pr(s|\mathbf{e})$ is a Poisson distribution with mean $\lambda(\mathbf{e})$, *i.e.*, that

$$\Pr(s|\mathbf{e}) = \frac{\lambda(\mathbf{e})^s e^{-\lambda(\mathbf{e})}}{s!}, \quad (7)$$

we can carry out the sum in Equation (1) explicitly,

$$\Pr(\mathbf{f}_1,\cdots,\mathbf{f}_m|\mathbf{e}) = e^{-\lambda(\mathbf{e})} \prod_{i=1}^{m} (e^{\lambda(\mathbf{e})p(\mathbf{f}_i|\mathbf{e})} - 1). \quad (8)$$

This is the form that we will assume throughout the remainder of the paper because of its simplicity. Notice that in this case, the probability of an entry is a product of factors, one for each of the translations that it contains.

The series $\mathbf{f}_1$, ..., $\mathbf{f}_m$ represents the translations of $\mathbf{e}$ that are included in the dictionary. We call this set of translations $D_\mathbf{e}$. Because we ignore everything about the dictionary except for these lists, a complete dictionary is just a collection of $D_\mathbf{e}$'s, one for each of the English phrases that has an entry. We treat each of these entries as independent and write the probability of the entire dictionary as

$$\Pr(D) \equiv \prod_{\mathbf{e}\in D} \Pr(D_\mathbf{e}|\mathbf{e}), \quad (9)$$

the product here running over all entries.

Equation (9) gives the probability of the dictionary in terms of the probabilities of the entries that make it up. The probabilities of these entries in turn are given by Equation (8) in terms of the probabilities, $p(\mathbf{f}|\mathbf{e})$, of individual French phrases given individual English phrases. Combining these two equations, we can write

$$\Pr(D) = \prod_{(\mathbf{e},\mathbf{f})\in D} (e^{\lambda(\mathbf{e})p(\mathbf{f}|\mathbf{e})} - 1) \prod_{\mathbf{e}\in D} e^{-\lambda(\mathbf{e})}. \quad (10)$$

We take $p(\mathbf{f}|\mathbf{e})$ to be given by the statistical model described in detail by Brown *et al.* [2]. Their model has a set of *translation probabilities*, $t(f|e)$, giving for each French word f and each English word e the probability that f will appear as (part of) a translation of e; a set of *fertility probabilities*, $n(\phi|e)$, giving for each integer ϕ and each English word e the probability that e will be translated as a phrase containing ϕ French words; and a set of distortion probabilities governing the placement of French words in the translation of an English phrase. They show how to estimate these parameters so as to maximize the probability,

$$\Pr(H) = \prod_{(\mathbf{e},\mathbf{f})\in H} p(\mathbf{f}|\mathbf{e}), \quad (11)$$

of a collection of pairs of aligned translations, $(\mathbf{e},\mathbf{f}) \in H$.

Let Θ represent the complete set of parameters of the model of Brown *et al.* [2], and let θ represent any one of the parameters. We extend the method of Brown *et al.* to develop a scheme for estimating Θ so as to maximize the joint probability of the corpus and the dictionary, $\Pr_\Theta(H,D)$. We assume that $\Pr_\Theta(H,D) = \Pr_\Theta(H)\Pr_\Theta(D)$. In general, it is possible only to find local maxima of $\Pr_\Theta(H,D)$ as a function of Θ, which we can do by applying the EM algorithm [3, 4]. The EM algorithm adjusts an initial estimate of Θ in a series of iterations. Each iteration consists of an estimation step in which a *count* is determined for each parameter, followed by a maximization step in which each parameter is replaced by a value proportional to its count. The count c_θ for a parameter θ is defined by

$$c_\theta = \theta \frac{\partial}{\partial\theta} \log \Pr_\Theta(H,D). \quad (12)$$

Because we assume that H and D are independent, we can write c_θ as the sum of a count for H and a count for D:

$$c_\theta = c_\theta(H) + c_\theta(D). \quad (13)$$

The corpus count is a sum of counts, one for each translation in the corpus. The dictionary count is also a sum of counts, but with each count weighted by a factor $\mu(\mathrm{e},\mathrm{f})$ which we call the *effective multiplicity* of the translation. Thus,

$$c_\theta(H) = \sum_{(\mathrm{e},\mathrm{f})\in H} c_\theta(\mathrm{e},\mathrm{f}) \tag{14}$$

and

$$c_\theta(D) = \sum_{(\mathrm{e},\mathrm{f})\in D} \mu(\mathrm{e},\mathrm{f}) c_\theta(\mathrm{e},\mathrm{f}) \tag{15}$$

with

$$c_\theta(\mathrm{e},\mathrm{f}) = \theta \frac{\partial}{\partial\theta} \log p_\Theta(\mathrm{f}|\mathrm{e}). \tag{16}$$

The effective multiplicity is just the expected number of times that our lexicographer observed the translation (e,f) given the dictionary and the corpus. In terms of the *a priori multiplicity*, $\mu_0(\mathrm{e},\mathrm{f}) = \lambda(\mathrm{e})p(\mathrm{f}|\mathrm{e})$, it is given by

$$\mu(\mathrm{e},\mathrm{f}) = \frac{\mu_0(\mathrm{e},\mathrm{f})}{1 - e^{-\mu_0(\mathrm{e},\mathrm{f})}}. \tag{17}$$

Figure 1 shows the effective multiplicity as a function of the *a priori* multiplicity. For small values of $\mu_0(\mathrm{e},\mathrm{f})$, $\mu(\mathrm{e},\mathrm{f})$ is approximately equal to $1 + \mu_0(\mathrm{e},\mathrm{f})/2$. For very large values, $\mu_0(\mathrm{e},\mathrm{f})$ and $\mu(\mathrm{e},\mathrm{f})$ are approximately equal. Thus, if we expect *a priori* that the lexicographer will see the translation (e,f) very many times, then the effective multiplicity will be nearly equal to this number, but even if we expect *a priori* that he will scarcely ever see a translation, the effective multiplicity for it cannot fall below 1. This is reasonable because in our model for the dictionary construction process, we assume that nothing can get into the dictionary unless it is seen at least once by the lexicographer.

RESULTS

We have used the algorithm described above to estimate translation probabilities and fertilities for our statistical model in two different ways. First, we estimated them from the corpus alone, then we estimated them from the corpus and the dictionary together. The corpus that we used is the proceedings of the Canadian Parliament described elsewhere [2]. The dictionary is a machine readable version of the HarperCollins Robert French Dictionary [1].

We do not expect that including information from the dictionary will have much effect on words that occur frequently in the corpus, and this is borne out by the data. But for words that are rare, we expect that there will be an effect.

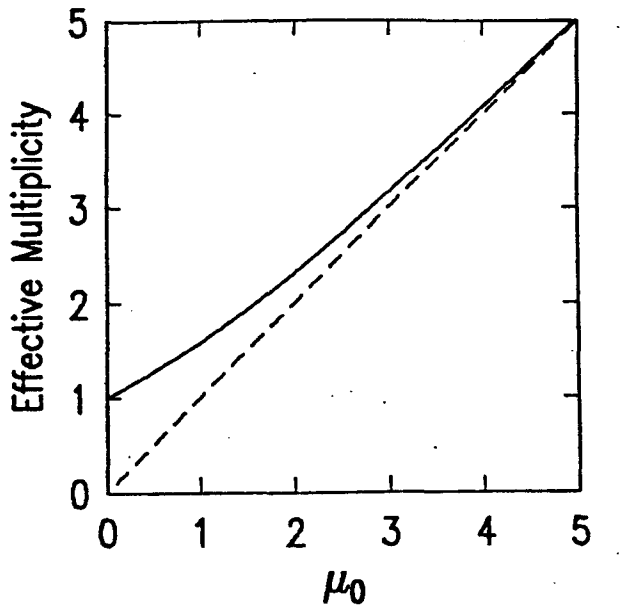

Figure 1: Effective multiplicity *vs* μ_0

f	$t(f\|e)$	ϕ	$n(\phi\|e)$
toundra	.233	3	.644
dans	.097	9	.160
autre	.048	1	.144
poser	.048	2	.021
ceux	.048	0	.029

Table 1: Parameters for *tundra*, corpus only

f	$t(f\|e)$	ϕ	$n(\phi\|e)$
toundra	.665	1	.855
dans	.040	3	.089
autre	.020	0	.029
poser	.020	9	.022
ceux	.020		

Table 2: Parameters for *tundra*, corpus and dictionary

Tables 1 and 2 show the two results for the English word *tundra*. The entry for *tundra* in the HarperCollins Robert French Dictionary [1] is simply the word *toundra*. We interpret this as a list with only one entry. We don't know how many times the lexicography ran across *tundra* translated as *toundra*, but we know that it was at least once, and we know that he never ran across it translated as anything else.

Even without the dictionary, *toundra* appears as the most probable translation, but with the dictionary, its probability is greatly improved. A more significant fact is the change in the fertility probabilities. Rare words have a tendency to act as garbage collectors in our system. This is why *tundra*, in the absence of guidance from the dictionary has, 3 as its most probable fertility and has a significant probability of fertility 9. With the dictionary, fertility 1 becomes the overwhelming favorite, and fertility 9 dwindles to insignificance.

Tables 3 and 4 show the trained parameters for *jungle*. The entry for *jungle* in the HarperCollins Robert French Dictionary is simply the word *jungle*. As with *tundra* using the dictionary enhances the probability of the dictionary translation of *jungle* and also improves the fertility substantially,

f	$t(f\|e)$	ϕ	$n(\phi\|e)$
jungle	.277	2	.401
dans	.072	1	.354
fouillis	.045	5	.120
domaine	.017	3	.080
devenir	.017	4	.020
imbroglio	.017	6	.019

Table 3: Parameters for *jungle*, corpus only

f	$t(f\|e)$	ϕ	$n(\phi\|e)$
jungle	.442	1	.598
dans	.057	5	.074
fouillis	.036	3	.049
domaine	.014	2	.024
devenir	.014	4	.012
imbroglio	.014	6	.012

Table 4: Parameters for *jungle*, corpus and dictionary

REFERENCES

[1] B. T. Atkins, A. Duval, R. C. Milne, P.-H. Cousin, H. M. A. Lewis, L. A. Sinclair, R. O. Birks, and M.-N. Lamy, *HarperCollins Robert French Dictionary*. New York: Harper & Row, 1990.

[2] P. F. Brown, S. A. DellaPietra, V. J. DellaPietra, and R. L. Mercer, "The mathematics of machine translation: Parameter estimation." Submitted to Computational Linguistics, 1992.

[3] L. Baum, "An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process," *Inequalities*, vol. 3, pp. 1–8, 1972.

[4] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. B, pp. 1–38, 1977.