A Statistical Approach to Sense Disambiguation in Machine Translation

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer

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

ABSTRACT

We describe a statistical technique for assigning senses to words. An instance of a word is assigned a sense by asking a question about the context in which the word appears. The question is constructed to have high mutual information with the word's translations.

INTRODUCTION

An alluring aspect of the statistical approach to machine translation rejuvenated by Brown, *et al.,* [1] is the systematic framework it provides for attacking the problem of lexical disambiguation. For example, the system they describe translates the French sentence *Je vais prendre la décision* as *I will make the decision*, thereby correctly interpreting *prendre* as *make*. The statistical translation model, which supplies English translations of French words, prefers the more common translation *take*, but the trigram language model recognizes that the three-word sequence *make the decision* is much more probable than *take the decision*.

The system is not always so successful. It incorrectly renders *Je vais prendre ma propre décision* as *I will take my own decision*. Here, the language model does not realize that *take my own decision* is improbable because *take* and *decision* no longer fall within a single trigram.

Errors such as this are common because our statistical models only capture local phenomena; if the context necessary to determine a translation falls outside the scope of our models, the word is likely to be translated incorrectly. However, if the relevant context is encoded locally, the word should be translated correctly. We can achieve this within the traditional paradigm of analysis – transfer – synthesis by incorporating into the analysis phase a sense-disambiguation component that assigns sense labels to French words. If *prendre* is labeled with one sense in the context of *décision* but with a different sense in other contexts, then the translation model will learn from training data that the first sense usually translates to *make*, whereas the other sense usually translates to *take*.

In this paper, we describe a statistical procedure for constructing a sense-disambiguation component that label words so as to elucidate their translations.

STATISTICAL TRANSLATION

As described by Brown, *et al.* [1], in the statistical approach to translation, one chooses for the translation of a French sentence F, that English sentence E which has the greatest probability, $\Pr(E|F)$, according to a model of the translation process. By Bayes' rule, $\Pr(E|F) = \Pr(E)\Pr(F|E)/\Pr(F)$. Since the denominator does not depend on E, the sentence for which $\Pr(E|F)$ is greatest is also the sentence for which the product $\Pr(E)\Pr(F|E)$ is greatest. The first term in this product is a statistical characterization of the English language and the second term is a statistical characterization of the process by which English sentences are translated into French. We can compute neither of these probabilities precisely. Rather, in statistical translation, we employ a *language model* $P_{model}(E)$ which provides an estimate of $\Pr(E)$ and a *translation model* which provides an estimate of $\Pr(F|E)$.

The performance of the system depends on the extent to which these statistical models approximate the actual probabilities. A useful gauge of this is the *cross entropy* [1]

$$H(\mathbf{E} \mid \mathbf{F}) \equiv - \sum_{E,F} \Pr(E,F) \log \mathrm{P}_{model}(E \mid F) \quad (1)$$

which measures the average uncertainty that the model has about the English translation E of a French sentence F. A better model has less uncertainty and thus a lower cross entropy.

A shortcoming of the architecture described above is that it requires the statistical models to deal directly with English and French sentences. Clearly the probability distributions $\Pr(E)$ and $\Pr(F \mid E)$ over sentences are immensely complicated. On the other hand, in practice the statistical models must be relatively simple in order that their parameters can be reliably estimated from a manageable amount of training data. This usually means that they are restricted to the modeling of local linguistic phenomena. As a result, the estimates $\mathrm{P}_{model}(E)$ and $\mathrm{P}_{model}(F \mid E)$ will be inaccurate.

This difficulty can be addressed by integrating statistical models into the traditional machine translation architecture of analysis-transfer-synthesis. The resulting system employs

1. An *analysis* component which encodes a French sentence F into an intermediate structure F'.
2. A *statistical transfer* component which translates F' a corresponding intermediate English structure E'. This component incorporates a language model, a translation model, and a decoder as before, but here these components deal with the intermediate structures rather than the sentences directly.
3. A *synthesis* component which reconstructs an English sentence E from E'.

For statistical modeling we require that the synthesis transformation $E' \mapsto E$ be invertible. Typically, analysis and synthesis will involve a sequence of successive transformations in which F' is incrementally constructed from F, or E is incrementally recovered from E'.

The purpose of analysis and synthesis is to facilitate the task of statistical transfer. This will be the case if the probability distribution $\Pr(E', F')$ is easier to model then the original distribution $\Pr(E, F)$. In practice this means that E' and F' should encode global linguistic facts about E and F in a local form.

The utility of the analysis and synthesis transformations can be measured in terms of cross-entropy. Thus transformations $F \rightarrow F'$ and $E' \rightarrow E$ are useful if we can construct models $\mathrm{P}'_{model}(F' \mid E')$ and $\mathrm{P}'_{model}(E')$ such that $H(\mathbf{E'} \mid \mathbf{F'}) < H(\mathbf{E} \mid \mathbf{F})$.

SENSE DISAMBIGUATION

In this paper we present a statistical method for automatically constructing analysis and synthesis transformations which perform *cross-lingual word-sense labeling.* The goal of such transformations is to label the words of a French sentence so as to elucidate their English translations, and, conversely, to label the words of an English sentence so as to elucidate their French translations. For example, in some contexts the French verb *prendre* translates as *to take*, but in other contexts it translates as *to make.* A sense disambiguation transformation, by examining the contexts, might label occurrences of *prendre* that likely mean *to take* with one label, and other occurrences of *prendre* with another label. Then the uncertainty in the translation of *prendre* given the label would be less than the uncertainty in the translation of *prendre* without the label. Although the label does not provide any information that is not already present in the context, it encodes this information locally. Thus a local statistical model for the transfer of labeled sentences should be more accurate than one for the transfer of unlabeled ones.

While the translation of a word depends on many words in its context, we can often obtain information by looking at only a single word. For example, in the sentence *Je vais prendre ma propre décision* (*I will make my own decision*), the verb *prendre* should be translated as *make* because its object is *décision.* If we replace *décision* by *voiture* then *prendre* should be translated as *take*: *Je vais prendre ma propre voiture* (*I will take my own car*). Thus we can reduce the uncertainity in the translation of *prendre* by asking a question about its object, which is often the first noun

[1] In this equation and in the remainder of the paper, we use bold face letters (e.g. **E**) for random variables and roman letters (e.g. E) for the values of random variables.

to its right, and we might assign a sense to *prendre* based upon the answer to this question.

In *Il doute que les nôtres gagnent* (*He doubts that we will win*), the word *il* should be translated as *he*. On the other hand, if we replace *doute* by *faut* then *il* should be translated as *it*: *Il faut que les nôtres gagnent* (*It is necessary that we win*). Here, we might assign a sense label to *il* by asking about the identity of the first verb to its right.

These examples motivate a sense-labeling scheme in which the label of a word is determined by a question about an *informant* word in its context. In the first example, the informant of *prendre* is the first noun to the right; in the second example, the informant of *il* is the first verb to the right. If we want to assign n senses to a word then we can consider a question with n answers.

We can fit this scheme into the framework of the previous section as follows:

The Intermediate Structures. The intermediate structures E' and F' consist of sequences of words labeled by their senses. Thus F' is a sentence over the expanded vocabulary whose 'words' f' are pairs (f, l) where f is a word in the original French vocabulary and l is its sense label. Similarly, E' is a sentence over the expanded vocabulary whose words e' are pairs (e, l) where e is an English word and l is its sense label.

The analysis and synthesis transformations. For each French word and each English word we choose an *informant site*, such as *first noun to the left*, and an n-ary question about the value of the informant at that site. The analysis transformation $F \mapsto F'$ and the inverse synthesis transformation $E \mapsto E'$ map a sentence to the intermediate structure in which each word is labeled by a sense determined by the question about its informant. The synthesis transformation $E' \mapsto E$ maps a labeled sentence to a sentence in which the labels have been removed.

The probability models. We use the translation model that was discussed in [1] for both $\mathrm{P}'_{model}(F' \mid E')$ and for $\mathrm{P}_{model}(F \mid E)$. We use a trigram language model [1] for $\mathrm{P}_{model}(E)$ and $\mathrm{P}'_{model}(E')$.

In order to construct these transformations we need to choose for each English and French word an informant and a question. As suggested in the previous section, a criterion for doing this is that of minimizing the cross entropy $H(\mathbf{E}' \mid \mathbf{F}')$. In the remainder of the paper we present an algorithm for doing this.

THE TRANSLATION MODEL

We begin by reviewing our statistical model for the translation of a sentence from one language to another [1]. In statistical French to English translation system we need to model transformations from English sentences E to French sentences F, or from intermediate English structures E' to intermediate French structures F'. However, it is clarifying to consider transformations from an arbitrary *source language* to an arbitrary *target language*.

Review of the Model

The purpose of a translation model is to compute the probability $\mathrm{P}_{model}(T \mid S)$ of transforming a source sentence S into a target sentence T. For our simple model, we assume that each word of S independently produces zero or more words from the target vocabulary and that these words are then ordered to produce T. We use the term *alignment* to refer to an association between words in T and words in S. The probability $\mathrm{P}_{model}(T \mid S)$ is the sum of the probabilities of all possible alignments A between S and T

$$\mathrm{P}_{model}(T \mid S) = \sum_{A} \mathrm{P}_{model}(T, A \mid S). \tag{2}$$

The joint probability $\mathrm{P}_{model}(T, A \mid S)$ of T and a particular alignment is given by

$$\mathrm{P}_{model}(T, A \mid S) = \prod_{t \in T} p(t \mid \hat{s}_A(t)) \prod_{s \in S} p(\hat{n}_A(s) \mid s) \, P_{distortion}(T, A \mid S). \tag{3}$$

Here $\hat{s}_A(t)$ is the word of S aligned with t in the alignment A, and $\hat{n}_A(s)$ is the number of words of T aligned with s in A. The *distortion model* $P_{distortion}$ describes the ordering of the words of T. We will not give it explicitly. The parameters in (3) are

1. The probabilities $p(n \mid s)$ that a word s in the source language generates n target words;
2. The probabilities $p(t \mid s)$ that s generates the word t;
3. The parameters of the distortion model.

We determine values for these parameters using *maximum likelihood training.* Thus we collect a large *bilingual corpus* consisting of pairs of sentences (S,T) which are translations of one another, and we seek parameter values that maximize the likelihood of this *training data* as computed by the model. This is equivalent to minimizing the cross entropy

$$H(\mathbf{T} \mid \mathbf{S}) = -\sum_{S,T} \mathrm{P}_{train}(S,T) \log \mathrm{P}_{model}(T \mid S) \quad (4)$$

where $\mathrm{P}_{train}(S,T)$ is the empirical distribution obtained by counting the number of times that the pair (S,T) occurs in the training corpus.

The Viterbi Approximation

The sum over alignments in (2) is too expensive to compute directly since the number of alignments increases exponentially with sentence length. It is useful to approximate this sum by the single term corresponding to the alignment, $\hat{A}(S,T)$, with greatest probability. We refer to this approximation as the *Viterbi approximation* and to $\hat{A}(S,T)$ as the *Viterbi alignment.*

Let $c(s,t)$ be the expected number of times that s is aligned with t in the Viterbi alignment of a pair of sentences drawn at random from the training data. Let $c(s,n)$ be the expected number of times that s is aligned with n words. Then

$$\begin{aligned} c(s,t) &= \sum_{S,T} \mathrm{P}_{train}(S,T)\, c(s,t \mid \hat{A}(S,T)) \\ c(s,n) &= \sum_{S,T} \mathrm{P}_{train}(S,T)\, c(s,n \mid \hat{A}(S,T)) \quad (5) \end{aligned}$$

where $c(s,t \mid A)$ is the number of times that s is aligned with t in the alignment A, and $c(s,n \mid A)$ is the number of times that s generates n target words in A. It can be shown [2] that these counts are also averages *with respect to the model*

$$\begin{aligned} c(s,t) &= \sum_{S,T} \mathrm{P}_{model}(S,T)\, c(s,t \mid \hat{A}(S,T)) \\ c(s,n) &= \sum_{S,T} \mathrm{P}_{model}(S,T)\, c(s,n \mid \hat{A}(S,T)). \quad (6) \end{aligned}$$

By normalizing the counts $c(s,t)$ and $c(s,n)$ we obtain probability distributions $p(s,t)$ and $p(s,n)$ [2]

$$p(s,t) = \frac{1}{norm} c(s,t) \qquad p(s,n) = \frac{1}{norm} c(s,n). \quad (7)$$

The conditional distributions $p(t \mid s)$ and $p(n \mid s)$ are the Viterbi approximation estimates for the parameters of the model. The marginals satisfy

$$\begin{aligned} \sum_n p(s,n) &= u(s) \qquad \sum_s p(s,t) = u(t) \\ \sum_t p(s,t) &= \frac{1}{norm} \bar{n}(s)\, u(s) \qquad (8) \end{aligned}$$

where $u(s)$ and $u(t)$ are the unigram distributions of s and t and $\bar{n}(s) = \sum_n p(n \mid s)\, n$ is the average number of target words aligned with s. These formulae reflect the fact that in any alignment each target word is aligned with exactly one source word.

CROSS ENTROPY

In this section we express the cross entropies $H(\mathbf{S} \mid \mathbf{T})$ and $H(\mathbf{S}' \mid \mathbf{T}')$ in terms of the information between source and target words.

In the Viterbi approximation the cross entropy $H(\mathbf{T} \mid \mathbf{S})$ is given by

$$H(\mathbf{T} \mid \mathbf{S}) = L_T \{ H(\mathbf{t} \mid \mathbf{s}) + H(\mathbf{n} \mid \mathbf{s}) \} \qquad (9)$$

where L_T is the average length of the target sentences in the training data, and $H(\mathbf{t} \mid \mathbf{s})$ and $H(\mathbf{n} \mid \mathbf{s})$ are the conditional entropies for the probability distributions $p(s,t)$ and $p(n,s)$:

$$\begin{aligned} H(\mathbf{t} \mid \mathbf{s}) &= -\sum_{s,t} p(s,t) \log p(t|s) \\ H(\mathbf{n} \mid \mathbf{s}) &= -\sum_{s,n} p(s,n) \log p(n|s). \qquad (10) \end{aligned}$$

We want a similar expression for the cross entropy $H(\mathbf{S} \mid \mathbf{T})$. Since

$$\mathrm{P}_{model}(S,T) = \mathrm{P}_{model}(T \mid S)\, \mathrm{P}_{model}(S),$$

this cross entropy depends on both the translation model, $\mathrm{P}_{model}(\mathbf{T} \mid \mathbf{S})$, and the language model, $\mathrm{P}_{model}(\mathbf{S})$. We now show that with a suitable additional approximation

$$H(\mathbf{S} \mid \mathbf{T}) = L_T \{ H(\mathbf{n} \mid \mathbf{s}) - I(\mathbf{s},\mathbf{t}) \} + H(\mathbf{S}) \quad (11)$$

[2] In these equations and in the remainder of the paper, we use the generic symbol $\frac{1}{norm}$ to denote a normalizing factor that converts counts to probabilities. We let the actual value of $\frac{1}{norm}$ be implicit from the context. Thus, for example, in the left hand equation of (7), the normalizing factor is $norm = \sum_{s,t} c(s,t)$ which equals the average length of target sentences. In the right hand equation of (7), the normalizing factor is the average length of source sentences.

where $H(\mathbf{S})$ is the cross entropy of $\mathrm{P}_{model}(S)$ and $I(\mathrm{s},\mathrm{t})$ is the mutual information between t and s for the probability distribution $p(s,t)$.

The additional approximation that we require is

$$H(\mathbf{T}) \approx L_T H(\mathrm{t}) \equiv -L_T \sum_t p(t) \log p(t) \qquad (12)$$

where $p(t)$ is the marginal of $p(s,t)$. This amounts to approximating $\mathrm{P}_{model}(T)$ by the unigram distribution that is closest to it in cross entropy. Granting this, formula (11) is a consequence of (9) and of the identities

$$\begin{aligned} H(\mathbf{S} \mid \mathbf{T}) &= H(\mathbf{T} \mid \mathbf{S}) - H(\mathbf{T}) + H(\mathbf{S}), \\ H(\mathrm{t}) &= H(\mathrm{t} \mid \mathrm{s}) + I(\mathrm{s},\mathrm{t}). \end{aligned} \qquad (13)$$

Next consider $H(\mathbf{S}' \mid \mathbf{T}')$. Let $S \rightarrow S'$ and $T \rightarrow T'$ be sense labeling transformations of the type discussed in Section 2. Assume that these transformations *preserve Viterbi alignments*; that is, if the words s and t are aligned in the Viterbi alignment for (S,T), then their sensed versions s' and t' are aligned in the Viterbi alignment for (S',T'). It follows that the word translation probabilities obtained from the Viterbi alignments satisfy $p(s,t) = \sum_{t' \in t} p(s,t') = \sum_{s' \in s} p(s',t)$ where the sums range over the sensed versions t' of t and the sensed versions s' of s.

By applying (11) to the cross entropies $H(\mathbf{S} \mid \mathbf{T})$, $H(\mathbf{S} \mid \mathbf{T}')$, and $H(\mathbf{S}' \mid \mathbf{T})$, it is not hard to verify that

$$\begin{aligned} H(\mathbf{S} \mid \mathbf{T}') &= H(\mathbf{S} \mid \mathbf{T}) - L_T \sum_t p(t) I(\mathrm{s},\mathrm{t}' \mid t) \\ H(\mathbf{S}' \mid \mathbf{T}) &= H(\mathbf{S} \mid \mathbf{T}) - \qquad\qquad (14) \\ &\qquad L_T \sum_s p(s)\{I(\mathrm{t},\mathrm{s}' \mid s) + I(\mathbf{n},\mathrm{s}', \mid s)\}. \end{aligned}$$

Here $I(\mathrm{s},\mathrm{t}' \mid t)$ is the conditional mutual information given a target word t between its translations s and its sensed versions t'; $I(\mathrm{t},\mathrm{s}' \mid s)$ is the conditional mutual information given a source word s between its translations t and its sensed versions s'; and $I(\mathbf{n},\mathrm{s}' \mid s)$ is the conditional mutual information given s between $\mathbf{n}$ and its sensed versions s'.

SELECTING QUESTIONS

We now present an algorithm for finding good informants and questions for sensing.

Target Questions

For sensing target sentences, a question about an informant is a function $\hat{c}$ from the target vocabulary into the set of possible senses. If the informant of t is x, then t is assigned the sense $\hat{c}(x)$. We want to choose the function $\hat{c}(x)$ to minimize the cross entropy $H(\mathbf{S} \mid \mathbf{T}')$. From formula (14), we see that this is equivalent to maximizing the conditional mutual information $I(\mathrm{s},\mathrm{t}' \mid t)$ between s and t'

$$I(\mathrm{s},\mathrm{t}' \mid t) = \sum_{s,x} p(s,x \mid t) \log \frac{p(s,\hat{c}(x) \mid t)}{p(s \mid t) p(\hat{c}(x) \mid t)} \qquad (15)$$

where $p(s,t,x)$ is the probability distribution obtained by counting the number of times in the Viterbi alignments that s is aligned with t and the value of the informant of t is x,

$$\begin{aligned} p(s,t,x) &= \frac{1}{norm} \sum_{S,T} \mathrm{P}_{train}(S,T)\, c(s,t,x \mid \hat{A}(S,T)\,) \\ p(s,t,c) &= \frac{1}{norm} \sum_{x:\hat{c}(x)=c} p(s,t,x). \end{aligned} \qquad (16)$$

An exhaustive search for the best $\hat{c}$ requires a computation that is exponential in the number of values of x and is not practical. In previous work [3] we found a good $\hat{c}$ using the flip-flop algorithm [4], which is only applicable if the number of senses is restricted to two. Since then, we have developed a different algorithm that can be used to find $\hat{c}$ for any number of senses. The algorithm uses the technique of *alternating minimization*, and is similar to the k-means algorithm for determining pattern clusters and to the generalized Lloyd algorithm for designing vector quantitizers. A discussion of alternating minimization, together with references, can be found in Chou [5].

The algorithm is based on the fact that, up to a constant independent of $\hat{c}$, the mutual information $I(\mathrm{s},\mathrm{t}' \mid t)$ can be expressed as an infimum over conditional probability distributions $q(\mathrm{s} \mid c)$,

$$\begin{aligned} &I(\mathrm{s},\mathrm{t}' \mid t) = \qquad\qquad (17) \\ &\quad \inf_q \sum_x p(x) D(p(\mathrm{s} \mid x,t)\ ;\ q(\mathrm{s} \mid \hat{c}(x)) + \text{constant} \end{aligned}$$

where

$$D(p(\mathrm{s})\ ;\ q(\mathrm{s})) \equiv \sum_s p(s) \log \frac{p(s)}{q(s)}. \qquad (18)$$

The best value of the information is thus an infimum over both the choice for $\hat{c}$ and the choice for the q. This suggests the following iterative procedure for obtaining a good $\hat{c}$:

1. For given q, find the best $\hat{c}$:
$$\hat{c}(x) = \text{argmin}_c D(p(\mathrm{s} \mid x, t) \; ; \; q(\mathrm{s} \mid c)).$$
2. For this $\hat{c}$, find the best q:
$$q(s \mid c) = \frac{1}{norm} \sum_{x:\hat{c}(x)=c} p(s, x \mid t).$$
3. Iterate steps (1) and (2) until no further increase in $I(\mathrm{s}, \mathrm{t}' \mid t)$ results.

Source Questions

For sensing source sentences, a question about an informant is a function $\hat{c}$ from the source vocabulary into the set of possible senses. We want to chose $\hat{c}$ to minimize the entropy $H(\mathbf{S}' \mid \mathbf{T})$. From (14) this is equivalent to maximizing the sum $I(\mathbf{t}, \mathrm{s}' \mid s) + I(\mathbf{n}, \mathrm{s}' \mid s)$. In analogy to (18),

$$\begin{aligned} I(\mathbf{t}, \mathrm{s}' \mid s) + I(\mathbf{n}, \mathrm{s}' \mid s) = & \qquad (19) \\ \inf_{q_1, q_2} \sum_x p(x) \quad & \{D(\mathrm{p}_{model}(\mathbf{t} \mid x, s) \; ; \; q_1(\mathbf{t} \mid \hat{c}(x)) \\ + \quad & D(\mathrm{p}_{model}(\mathbf{n} \mid x, s) \; ; \; q_2(\mathbf{n} \mid \hat{c}(x))\}. \end{aligned}$$

and we can again find a good $\hat{c}$ by alternating minimization.

CONCLUSION

In this paper we presented a general framework for integrating analysis and synthesis with statistical translation, and within this framework we investigated cross-lingual sense labeling. We gave an algorithm for automatically constructing a simple labeling transformation that assigns a sense to a word by asking a question about a single word of the context. In a companion paper [3] we present results of translation experiments using a sense-labeling component that employs a similar algorithm. We are currently studying the automatic construction of more complex transformations which utilize more detailed contextual information.

REFERENCES

[1] P. Brown, J. Cocke, S. DellaPietra, V. DellaPietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, pp. 79–85, June 1990.

[2] P. Brown, S. DellaPietra, V. DellaPietra, and R. Mercer, "Initial estimates of word translation probabilities." In preparation.

[3] P. Brown, S. DellaPietra, V. DellaPietra, and R. Mercer, "Word sense disambiguation using statistical methods," in *Proceedings 29th Annual Meeting of the Association for Computational Linguistics*, (Berkeley, CA), June 1991.

[4] A. Nadas, D. Nahamoo, M. Picheny, and J. Powell, "An iterative "flip-flop"" approximation of the most informative split in the construction of decision trees," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (Toronto, Canada), May 1991.

[5] P. Chou, *Applications of Information Theory to Pattern Recognition and the Design of Decision Trees and Trellises.* PhD thesis, Stanford University, June 1988.